The Annals of Statistics
2011, Vol. 39, No. 5, 2686–2715
DOI: 10.1214/11-AOS914

ROBUST RECOVERY OF MULTIPLE SUBSPACES BY GEOMETRIC L_P MINIMIZATION[1]

By Gilad Lerman and Teng Zhang

University of Minnesota

We assume i.i.d. data sampled from a mixture distribution with K components along fixed d-dimensional linear subspaces and an additional outlier component. For $p > 0$, we study the simultaneous recovery of the K fixed subspaces by minimizing the l_p-averaged distances of the sampled data points from any K subspaces. Under some conditions, we show that if $0 < p \leq 1$, then all underlying subspaces can be precisely recovered by l_p minimization with overwhelming probability. On the other hand, if $K > 1$ and $p > 1$, then the underlying subspaces cannot be recovered or even nearly recovered by l_p minimization. The results of this paper partially explain the successes and failures of the basic approach of l_p energy minimization for modeling data by multiple subspaces.

1. Introduction.

In the last decade, many algorithms have been developed to model data by multiple subspaces. Such hybrid linear modeling (HLM) was motivated by concrete problems in computer vision as well as by nonlinear dimensionality reduction. HLM is the simplest geometric framework for nonlinear dimensionality reduction. Nevertheless, very little theory has been developed to justify the performance of existing methods. Here we give a rigorous analysis of the recovery of multiple subspaces via an energy minimization.

One can model a data set $\mathcal{X}$ with K subspaces obtained by minimizing the following energy over the subspaces $\mathrm{L}_1, \ldots, \mathrm{L}_K$:

$$e_{l_p}(\mathcal{X}, \mathrm{L}_1, \ldots, \mathrm{L}_K) = \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{dist}^p\left(\mathbf{x}, \bigcup_{i=1}^{K} \mathrm{L}_i\right), \tag{1}$$

Received April 2011.

[1] Supported in part by NSF Grants DMS-09-15064 and DMS-09-56072.

AMS 2000 subject classifications. 62H30, 62G35, 68Q32.

Key words and phrases. Detection, clustering, multiple subspaces, hybrid linear modeling, optimization on the Grassmannian, robustness, geometric probability, high-dimensional data.

where $\operatorname{dist}(\cdot,\cdot)$ denotes the Euclidean distance and $p > 0$ is a fixed parameter. For simplicity, we assume that $L_1, \ldots, L_K$ are linear subspaces of the same dimension d, and we refer to them as d-subspaces (generalizations are discussed in Sections 5.6 and 5.7). We also assume that the data set $\mathcal{X}$ contains i.i.d. samples from a mixture distribution μ with K components along fixed d-subspaces and an additional outlier component. The recovery problem asks whether with overwhelming probability the minimization of (1) recovers the underlying subspaces of μ. We show here that when $p \leq 1$ the answer to this problem is positive, whereas when $p > 1$ it is negative.

Recovery problems are common in statistics, for example, recovering a single subspace in least squares type problems or recovering multiple centers as in K-means. However, our recent setting requires novel developments. One issue is the strong geometric nature of our problem, resulting from an optimization on a product space of Grassmannians. The other is the difficulty of approximating the problem by convex optimization (as we clarify in Section 5.1). Thus, even though it is an elementary problem in statistical learning, it requires the development of techniques which are currently not widely common in statistics.

1.1. *Background and related work.* Many algorithms have been developed for HLM (see, e.g., [1, 5, 8–11, 13, 14, 20–26]), and they find diverse applications in several areas, such as motion segmentation in computer vision, hybrid linear representation of images, classification of face images and temporal segmentation of video sequences (see, e.g., [14, 23, 26]). HLM is the simplest nonlinear data modeling and fits within the broader frameworks of modeling data by mixture of manifolds [3] and by Whitney's stratified space [4].

The K-subspaces algorithm [5, 10, 22] is the most basic heuristic for HLM, and it suggests an iterative procedure attempting to minimize the energy (1) with $p = 2$. It generalizes the K-means algorithm, which models data by K centers, that is, 0-dimensional affine subspaces. Numerical experiments by Zhang et al. [25] have shown that the K-subspaces algorithm is in general not robust to outliers, whereas a different method aiming to minimize (1) with $p = 1$ seems to be robust to outliers.

There has been little investigation into performance guarantees of the various HLM algorithms. Nevertheless, the accuracy of segmentation under some sampling assumptions was analyzed for two spectral-type HLM algorithms in [7] and [3], where [3] also quantified the tolerance to outliers ([3] considers only the asymptotic case, though applies to modeling by multiple manifolds). For the K-means algorithm (which only applies to 0-dimensional affine subspaces), Pollard has established strong consistency [16] and a central limit theorem [17].

In [12], we analyzed the l_p-recovery of the "most significant" subspace among multiple subspaces and outliers with spherically symmetric underlying distributions. We assume here a similar (though weaker) underlying model and rely on some of the estimates already developed there.

1.2. *Basic conventions and notation.* We denote by $\mathrm{G}(D,d)$ the Grassmannian, that is, the manifold of d-subspaces of $\mathbb{R}^D$. We measure distances between F and G in $\mathrm{G}(D,d)$ by the metric

$$\mathrm{dist}_{\mathrm{G}}(\mathrm{F},\mathrm{G}) = \sqrt{\sum_{i=1}^{d} \theta_i^2}, \tag{2}$$

where $\{\theta_i\}_{i=1}^d$ are the principal angles between F and G. We use this distance since there is a simple formula for the geodesic lines on the Grassmannian equipped with this distance (see, e.g., [12], equation 12), which is applied in this paper. We distinguish elements in the K-fold product space $\mathrm{G}(D,d)^K$ by the l_∞ norm, that is,

$$\mathrm{dist}_{\mathrm{G}^K}((\mathrm{L}_1,\dots,\mathrm{L}_K),(\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)) = \max_{i=1,\dots,K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i,\hat{\mathrm{L}}_i)). \tag{3}$$

Following [15], Section 3.9, we denote by $\gamma_{D,d}$ the "uniform" distribution on $\mathrm{G}(D,d)$.

We denote by $a \vee b$ and $a \wedge b$ the maximum and minimum of a and b, respectively. We designate the support of a distribution μ by $\mathrm{supp}(\mu)$. By saying "with overwhelming probability" or, in short, "w.o.p.," we mean that the underlying probability is at least $1 - Ce^{-N/C}$, where C is a constant independent of N.

1.3. *Setting of this paper.* We assume an i.i.d. data set $\mathcal{X} \subseteq \mathbb{R}^D$ of size N sampled from a mixture distribution representing a hybrid linear model around distinct d-subspaces, $\{\mathrm{L}_i^*\}_{i=1}^K$. We in fact consider two different types of models, but both of them have the same basic structure.

We assume K distributions, μ_i, each supported on a corresponding and distinct d-subspace, L_i^*, a noise level $\varepsilon \geq 0$, and an outlier distribution, denoted by μ_0. Furthermore, for each $1 \leq i \leq K$ we have a distinct noise distribution $\nu_{i,\varepsilon}$ with bounded support in the orthogonal complement L_i^*. We assume that the pth moments of $\{\|\nu_{i,\varepsilon}\|\}_{i=1}^K$ are smaller than ε^p for all $0 < p \leq 1$ ($p < 1$ is only needed when we consider l_p minimization with $p < 1$). Moreover, if $\varepsilon = 0$, then $\{\nu_{i,0}\}_{i=1}^K$ are the Dirac δ distributions supported on the origin within the corresponding subspaces orthogonal to $\{\mathrm{L}_i^*\}_{i=1}^K$.

We assume that the underlying distributions, $\{\mu_i\}_{i=0}^K$, have bounded supports (or possibly sub-Gaussian as explained in Section 5.3). In order to simplify our estimates, we further assume that $\mathrm{supp}(\mu_i) \subseteq \mathrm{B}(\mathbf{0},1)$ for $0 \leq i \leq K$.

From these pieces we construct the mixture distribution μ_ε,

$$\mu_\varepsilon = \alpha_0\mu_0 + \sum_{i=1}^{K} \alpha_i \mu_i \times \nu_{i,\varepsilon}, \tag{4}$$

where $\alpha_0 \geq 0$, $\alpha_i > 0$ $\forall 1 \leq i \leq K$ and $\sum_{i=0}^{K} \alpha_i = 1$. If $\varepsilon = 0$, then for convenience we replace the notation μ_ε by μ, that is,

$$\mu = \alpha_0\mu_0 + \sum_{i=1}^{K} \alpha_i \mu_i. \tag{5}$$

Within this basic framework, we analyze two different models. For $\varepsilon \geq 0$ and μ_ε as in (4), we say that μ_ε is a *weakly spherically symmetric HLM distribution with noise level* ε if the $\{\mu_i\}_{i=1}^K$ are generated by rotations (in $\mathbb{R}^D$) of a single distribution $\hat{\mu}$, such that $\hat{\mu}(\{\mathbf{0}\}) < 1$, $\operatorname{supp}(\hat{\mu}) \subseteq \mathrm{B}(\mathbf{0}, 1) \cap \hat{\mathrm{L}}$ for some d-subspace $\hat{\mathrm{L}} \subset \mathbb{R}^D$ and $\hat{\mu}$ is spherically symmetric within $\hat{\mathrm{L}}$ (i.e., invariant to rotations within $\hat{\mathrm{L}}$).

Our second model has weaker assumptions on the distributions of inliers and a slightly stronger assumption on the distribution of outliers. For $\varepsilon \geq 0$ and μ_ε as in (4), we say that μ_ε is a *weak HLM distribution with noise level* ε if $\mu_i(\{\mathbf{0}\}) < 1$ $\forall 1 \leq i \leq K$, $\operatorname{supp}(\mu_\varepsilon) \subseteq \mathrm{B}(\mathbf{0}, 1)$ and for some $r > 0$ the uniform distribution on $\mathrm{B}(\mathbf{0}, r)$ is absolutely continuous w.r.t. the restriction of μ_0 to $\mathrm{B}(\mathbf{0}, r)$.

Our theory uses the constant $\tau_0 \equiv \tau_0(d, p, \{\mu_i\}_{i=1}^K)$. We delay its definition to the proofs [see (11)], but use it in the formulation of Theorems 1.1 and 1.2.

1.4. *Statistical problems of this paper.* We address here two statistical problems. The simpler one is implicit in this introduction, though clear from the proofs. It asks whether the underlying subspaces $\{\mathrm{L}_i^*\}_{i=1}^K$ can be recovered when $\varepsilon = 0$ by minimizing $\mathbb{E}_\mu(\operatorname{dist}^p(\mathbf{x}, \bigcup_{i=1}^K \mathrm{L}_i))$ over $\{\mathrm{L}_i\}_{i=1}^K \subset \mathrm{G}(D, d)$. The main problem can be formulated using the empirical distribution μ_N of i.i.d. sample of size N from μ. It asks whether $\{\mathrm{L}_i^*\}_{i=1}^K$ can be recovered (w.o.p.) by minimizing $\mathbb{E}_{\mu_N}(\operatorname{dist}^p(\mathbf{x}, \bigcup_{i=1}^K \mathrm{L}_i))$, which is equivalent to minimizing (1). In the noisy case, we extend these problems to near recovery. When $K > 1$ and $d \geq 1$, these problems are nontrivial and require complicated geometric estimates.

1.5. *Main theory.* We first formulate the exact recovery of $\{\mathrm{L}_i^*\}_{i=1}^K$ as the unique global minimizer of the l_p energy (1) when $0 < p \leq 1$.

THEOREM 1.1. *Assume that μ is a weakly spherically symmetric HLM distribution on $\mathbb{R}^D$ without noise ($\varepsilon = 0$) and with underlying subspaces*

$\{\mathrm{L}_i^*\}_{i=1}^K \subseteq \mathbb{R}^D$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$. Let $\mathcal{X}$ be an i.i.d. data set sampled from μ. If $0 < p \leq 1$ and

$$\alpha_0 < \tau_0 \cdot \min_{i=1,\dots,K} \alpha_i \cdot \left(1 \wedge \min_{1\leq i,j\leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)^p/2^p\right), \tag{6}$$

then w.o.p. the set $\{\mathrm{L}_1^*, \dots, \mathrm{L}_K^*\}$ is the unique global minimizer of the energy (1) among all d-subspaces in $\mathbb{R}^D$.

Theorem 1.1 extends to the noisy case by allowing near-recovery as follows (a counterexample for asymptotic exact recovery is shown in Section 3.2).

THEOREM 1.2. *Assume that $\varepsilon > 0$ and μ_ε is a weakly spherically symmetric HLM distribution of noise level ε on $\mathbb{R}^D$ with K d-subspaces $\{\mathrm{L}_i^*\}_{i=1}^K \subseteq \mathbb{R}^D$ and mixture coefficients $\{\alpha_i\}_{i=0}^K$. Let $\mathcal{X}$ be an i.i.d. data sampled from μ_ε. If $0 < p \leq 1$ and*

$$\varepsilon < 3^{-1/p}\left(\tau_0 \cdot \min_{i=1,\dots,K} \alpha_i \cdot \left(1 \wedge \min_{1\leq i,j\leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)^p/2^p\right) - \alpha_0\right)^{1/p}, \tag{7}$$

then any minimizer of (1) in $\mathrm{G}(D,d)^K$ has a distance smaller than

$$f \equiv f(\varepsilon, K, d, p, \{\alpha_i\}_{i=1}^K) = 3^{1/p} \cdot \left(\tau_0 \min_{1\leq j\leq K} \alpha_j - \alpha_0\right)^{-1/p} \cdot \varepsilon \tag{8}$$

from one of the permutations of $(\mathrm{L}_1^*, \dots, \mathrm{L}_K^*)$ with overwhelming probability.

At last, we formulate the impossibility to recover $\{\mathrm{L}_i^*\}_{i=1}^K$ by l_p minimization when $p > 1$ (the constants δ_0 and κ_0 in our formulation are estimated in Section 4.5.5).

THEOREM 1.3. *Assume an i.i.d. sample of K d-subspaces $\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d)$ from the "uniform" distribution on $\mathrm{G}(D,d)$, $\gamma_{D,d}$. For $\varepsilon \geq 0$ and the sample $\{\mathrm{L}_i^*\}_{i=1}^K$, let μ_ε be a weak HLM distribution with noise level ε and let $\mathcal{X}$ be an i.i.d. data set of size N sampled from μ_ε. If $p > 1$ and $K > 1$, then for almost every $\{\mathrm{L}_i^*\}_{i=1}^K$ (w.r.t. $\gamma_{D,d}^K$) there exist positive constants δ_0 and κ_0, independent of N, such that for any $\varepsilon < \delta_0$ the minimizer of (1), $\hat{\mathrm{L}}_1, \dots, \hat{\mathrm{L}}_K$, satisfies w.o.p.:*

$$\mathrm{dist}_{\mathrm{G}^K}((\hat{\mathrm{L}}_1, \dots, \hat{\mathrm{L}}_K), (\mathrm{L}_1^*, \dots, \mathrm{L}_K^*)) > \kappa_0. \tag{9}$$

The above theorems have direct implications for HLM with spherically symmetric sampling along the subspaces. Theorems 1.1 and 1.2 clarify to some extent the robustness of two recent algorithms for HLM, which use the l_1 energy (1): Median K-Flats (MKF) [25] and Local Best-fit Flats (LBF) [27]. Theorem 1.3 explains why common HLM strategies that use the l_2 energy (1) (e.g., K-subspaces) are generally not robust to outliers.

1.6. *Structure of the paper.* Theorems 1.1, 1.2 and 1.3 are proved in Sections 2, 3 and 4, respectively. Section 5 discusses possible extensions as well as limitations of our theory and suggests some open directions.

2. Proof of Theorem 1.1.

2.1. *Preliminaries.* We view the energy $e_{l_p}(\mathcal{X}, \mathrm{L}_1, \ldots, \mathrm{L}_K)$ as a function defined on $\mathrm{G}(D,d)^K$ while being conditioned on the fixed data set $\mathcal{X}$. Therefore, the minimizer of $e_{l_p}(\mathcal{X}, \mathrm{L}_1, \ldots, \mathrm{L}_K)$ is an element $(\mathrm{L}'_1, \ldots, \mathrm{L}'_K)$ in $\mathrm{G}(D,d)^K$. Since any permutation of its K coordinates in $\mathrm{G}(D,d)$ results in another minimizer, we sometimes say that the set $\{\mathrm{L}'_1, \ldots, \mathrm{L}'_K\}$ is a minimizer [instead of $(\mathrm{L}'_1, \ldots, \mathrm{L}'_K)$].

We denote $e_{l_p}(\mathbf{x}, \mathrm{L}_1, \ldots, \mathrm{L}_K) := e_{l_p}(\{\mathbf{x}\}, \mathrm{L}_1, \ldots, \mathrm{L}_K)$ and view it as a function on $\mathbb{R}^D \times \mathrm{G}(D,d)^K$.

We denote the set of all permutations of $(1, 2, \ldots, K)$ by $\mathcal{P}_K$. We designate an open ball in $\mathrm{G}(D,d)$ by $\mathrm{B}_{\mathrm{G}}(\mathrm{L}, r)$ as opposed to the Euclidean open ball in $\mathbb{R}^D$, $\mathrm{B}(\mathbf{x}, r)$.

We partition $\mathcal{X}$ into the subsets $\{\mathcal{X}_i\}_{i=0}^K$ with $\{N_i\}_{i=0}^K$ points sampled according to the distributions $\{\mu_i\}_{i=0}^K$.

We define

$$\psi_{\mu_1}(t) = \mu_1(\mathbf{x} \in \mathbb{R}^D : -t < |\mathbf{x}^T \mathbf{v}| < t), \tag{10}$$

where $\mathbf{v}$ is an arbitrarily fixed unit vector in L_1^* [due to the spherical symmetry of μ_1 within L_1^*, (10) is independent of $\mathbf{v}$]. We note that since $\{\mu_i\}_{i=1}^K$ are generated by a single distribution, $\psi_{\mu_1}(t) = \psi_{\mu_i}(t)\ \forall 2 \le i \le K$. The invertibility of ψ_{μ_1} is established in [12], Appendix A.2, and an estimate of ψ_{μ_1} for a uniform distribution on a d-dimensional ball appears in [12], Appendix A.1.

Theorem 1.1 uses the constant τ_0, which we can now define as follows:

$$\tau_0 := \frac{(1 - \mu_1(\{\mathbf{0}\})) \cdot 2^{p-1} \cdot \psi_{\mu_1}^{-1}((1 + (2K-1)\mu_1(\{\mathbf{0}\}))/(2K))^p}{(\pi\sqrt{d})^p}. \tag{11}$$

In the special case where μ_1 is the uniform distribution on $\mathrm{B}(\underline{0}, 1) \cap \mathrm{L}_1$, then the estimate of ψ_μ in [12], Section A.1, implies the following lower bound for τ_0:

$$\tau_0 > \frac{1}{2^{p+1} \cdot K^p \cdot d^{3p/2}}.$$

Consequently, Theorem 1.1 holds in this case if τ_0 in (6) is replaced by $1/(2^{p+1} \cdot K^p \cdot d^{3p/2})$. Furthermore, it follows from basic scaling arguments that if μ_1 is the uniform distribution on $\mathrm{B}(\underline{0}, r_1) \cap \mathrm{L}_1$ and $\operatorname{supp}(\mu_0) \subseteq \mathrm{B}(\underline{0}, r_2)$, where r_1 and r_2 are any positive numbers, then

$$\tau_0 > \frac{r_1^p}{2^{p+1} \cdot K^p \cdot d^{3p/2} \cdot r_2^p}.$$

2.2. *Auxiliary lemmata.* The following lemmata are used throughout this proof (Lemma 2.1 is proved in the Appendix and Lemma 2.2 in [12], Appendix A.2).

LEMMA 2.1. *Suppose that* $\mathrm{L}_1, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K \in \mathrm{G}(D,d), p > 0$ *and* μ_1 *is a spherically symmetric distribution in* $\mathrm{B}(\mathbf{0},1) \cap \mathrm{L}_1$. *If* $\min_{1 \le j \le K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \hat{\mathrm{L}}_j) > \varepsilon$, *then*

$$\mathbb{E}_{\mu_1}(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K)) > \tau_0 \varepsilon^p.$$

LEMMA 2.2. *For any* $\mathbf{x} \in \mathbb{R}^D$ *and* $\mathrm{L}_1, \mathrm{L}_2 \in \mathrm{G}(D,d)$,

$$|\mathrm{dist}(\mathbf{x}, \mathrm{L}_1) - \mathrm{dist}(\mathbf{x}, \mathrm{L}_2)| \le \|\mathbf{x}\| \, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2).$$

2.3. *Proof in expectation.* We verify Theorem 1.1 "in expectation," whereas later sections extend the proof to hold w.o.p. We use the following notation w.r.t. the fixed d-subspaces $\mathrm{L}_1^*, \mathrm{L}_2^*, \ldots, \mathrm{L}_K^*, \hat{\mathrm{L}}_1, \hat{\mathrm{L}}_2, \ldots, \hat{\mathrm{L}}_K \in \mathrm{G}(D,d)$:

$$I(i) = \underset{1 \le j \le K}{\arg\min}\, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \hat{\mathrm{L}}_j) \qquad \forall 1 \le i \le K \tag{12}$$

and

$$d_0 = \min_{i_1, i_2, \ldots, i_K \in \mathcal{P}_K} \mathrm{dist}_{\mathrm{G}^K}((\mathrm{L}_{i_1}^*, \ldots, \mathrm{L}_{i_K}^*), (\hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K)). \tag{13}$$

The "expected version" of Theorem 1.1 is formulated and proved as follows.

PROPOSITION 2.1. *Suppose that* $\hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K$ *are arbitrary subspaces in* $\mathrm{G}(D,d)$, $0 < p \le 1$, *and* I *is defined w.r.t.* $\{\hat{\mathrm{L}}_i\}_{i=1}^K$ *and the underlying subspaces* $\{\mathrm{L}_i^*\}_{i=1}^K$. *If* $(I(1), \ldots, I(K))$ *is a permutation of* $(1, \ldots, K)$, *then*

$$\begin{aligned} &\mathbb{E}_\mu e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K) - \mathbb{E}_\mu e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*) \\ &\quad \ge \Big(\tau_0 \min_{1 \le j \le K} \alpha_j - \alpha_0\Big) d_0^p. \end{aligned} \tag{14}$$

On the other hand, if $(I(1), \ldots, I(K))$ *is not a permutation of* $(1, \ldots, K)$, *then*

$$\begin{aligned} &\mathbb{E}_\mu e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K) - \mathbb{E}_\mu e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*) \\ &\quad \ge \tau_0 \Big(\min_{1 \le j \le K} \alpha_j\Big) \Big(\min_{1 \le i,j \le K} \mathrm{dist}_{\mathrm{G}}^p(\mathrm{L}_i^*, \mathrm{L}_j^*)/2\Big) - \alpha_0. \end{aligned} \tag{15}$$

PROOF. We define

$$M = \underset{1 \le i \le K}{\arg\max}\, \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \hat{\mathrm{L}}_{I(i)}).$$

Assume first that $(I(1),\dots,I(K))$ is a permutation of $(1,\dots,K)$. Using the definition of I, we have

$$\begin{aligned}\min_{1\le j\le K}\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_M^*,\hat{\mathrm{L}}_j) &= \operatorname{dist}_{\mathrm{G}}(\mathrm{L}_M^*,\hat{\mathrm{L}}_{I(M)})\\ &= \operatorname{dist}_{\mathrm{G}^K}((\mathrm{L}_1^*,\dots,\mathrm{L}_K^*),(\hat{\mathrm{L}}_{I(1)},\dots,\hat{\mathrm{L}}_{I(K)}))\\ &= d_0.\end{aligned}\tag{16}$$

Combining (16) with Lemma 2.1, we obtain that

$$\begin{aligned}&\mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)-\mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x},\mathrm{L}_1^*,\dots,\mathrm{L}_K^*)\\ &\qquad= \mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K) > \tau_0 d_0^p.\end{aligned}\tag{17}$$

For any $\mathbf{x}\in\mathcal{X}_0$, let $m(\mathbf{x}) = \arg\min_{1\le i\le K}\operatorname{dist}(\mathbf{x},\mathrm{L}_i^*)$, $\hat{m}(\mathbf{x}) = \arg\min_{1\le i\le K}\operatorname{dist}(\mathbf{x},\hat{\mathrm{L}}_i)$ and note that

$$\begin{aligned}&e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)-e_{l_p}(\mathbf{x},\mathrm{L}_1^*,\dots,\mathrm{L}_K^*)\\ &\qquad= \operatorname{dist}(\mathbf{x},\hat{\mathrm{L}}_{\hat{m}(\mathbf{x})})^p-\operatorname{dist}(\mathbf{x},\mathrm{L}^*_{m(\mathbf{x})})^p\\ &\qquad\ge \operatorname{dist}(\mathbf{x},\hat{\mathrm{L}}_{\hat{m}(\mathbf{x})})^p-\operatorname{dist}(\mathbf{x},\mathrm{L}^*_{I^{-1}(\hat{m}(\mathbf{x}))})^p\\ &\qquad\ge -\|\mathbf{x}\|^p\operatorname{dist}_{\mathrm{G}}(\hat{\mathrm{L}}_{\hat{m}(\mathbf{x})},\mathrm{L}^*_{I^{-1}(\hat{m}(\mathbf{x}))})^p\\ &\qquad\ge -\|\mathbf{x}\|^p d_0^p\ge -d_0^p,\end{aligned}\tag{18}$$

where the second inequality in (18) uses Lemma 2.2. Therefore,

$$\mathbb{E}_{\mu_0} e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)-\mathbb{E}_{\mu_0} e_{l_p}(\mathbf{x},\mathrm{L}_1^*,\dots,\mathrm{L}_K^*) > -d_0^p.\tag{19}$$

At last, we observe that

$$\begin{aligned}&\mathbb{E}_{\mu} e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)-\mathbb{E}_{\mu} e_{l_p}(\mathbf{x},\mathrm{L}_1^*,\dots,\mathrm{L}_K^*)\\ &\quad\ge \alpha_M(\mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)-\mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x},\mathrm{L}_1^*,\dots,\mathrm{L}_K^*))\\ &\qquad+\alpha_0(\mathbb{E}_{\mu_0} e_{l_p}(\mathbf{x},\hat{\mathrm{L}}_1,\dots,\hat{\mathrm{L}}_K)-\mathbb{E}_{\mu_0} e_{l_p}(\mathbf{x},\mathrm{L}_1^*,\dots,\mathrm{L}_K^*)).\end{aligned}\tag{20}$$

The proposition in this case thus follows from (17), (19) and (20).

Next, we assume that $I(1),\dots,I(K)$ is not a permutation of $1,2,\dots,K$. In this case, there exist $1\le n_1,n_2\le K$ such that $I(n_1)=I(n_2)$ and, consequently,

$$\begin{aligned}2\min_{1\le j\le K}\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_M^*,\hat{\mathrm{L}}_j) &= 2\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_M^*,\hat{\mathrm{L}}_{I(M)})\\ &\ge \operatorname{dist}_{\mathrm{G}}(\mathrm{L}_{n_1}^*,\hat{\mathrm{L}}_{I(n_1)})+\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_{n_2}^*,\hat{\mathrm{L}}_{I(n_2)})\\ &\ge \operatorname{dist}_{\mathrm{G}}(\mathrm{L}_{n_1}^*,\mathrm{L}_{n_2}^*)\\ &\ge \min_{1\le i,j\le K}\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_i^*,\mathrm{L}_j^*).\end{aligned}\tag{21}$$

Combining (21) and Lemma 2.1 [applied with $\varepsilon = \min_{1\leq i,j\leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)/2$], we obtain that

$$\begin{aligned}(22)\qquad &\mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K) - \mathbb{E}_{\mu_M} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*)\\ &\quad > \tau_0\Big(\min_{1\leq i,j\leq K} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)/2\Big)^p.\end{aligned}$$

Finally, since the support of μ_0 is contained in $\mathrm{B}(\mathbf{0}, 1)$, we note that

$$(23)\qquad \mathbb{E}_{\mu_0} e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K) - \mathbb{E}_{\mu_0} e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*) \geq -1.$$

The proposition is thus concluded from (20), (22) and (23). □

2.4. *Proof in a local ball by calculus on the Grassmannian.* We cannot directly extend (14) to an estimate w.o.p., since its lower bound is a multiplication of d_0^p, which approaches zero as the set $\{\mathrm{L}_i\}_{i=1}^K$ approaches $\{\mathrm{L}_i^*\}_{i=1}^K$. We will need to exclude a ball in $\mathrm{G}(D,d)^K$ around $\{\mathrm{L}_i^*\}_{i=1}^K$ before such an extension. We thus prove here that $\{\mathrm{L}_i^*\}_{i=1}^K$ is a unique global minimizer w.o.p. in a local ball. In Section 2.5 we extend Proposition 2.1 to an estimate w.o.p. outside this ball and conclude the theorem.

We show that there exists a sufficiently small number γ_1 such that $\{\mathrm{L}_i^*\}_{i=1}^K$ is the unique global minimizer w.o.p. of e_{l_p} in $\mathrm{B}_{\mathrm{G}}((\mathrm{L}_{i_1}^*, \ldots, \mathrm{L}_{i_K}^*), \gamma_1)$. Since e_{l_p} is permutation invariant, it is also the unique global minimizer in

$$\bigcup_{i_1,i_2,\ldots,i_K\in\mathcal{P}_K} \mathrm{B}_{\mathrm{G}}((\mathrm{L}_{i_1}^*, \ldots, \mathrm{L}_{i_K}^*), \gamma_1).$$

In order to simplify notation in this part of the proof, we will adopt WLOG the convention that the RHS of (3) occurs at $i=1$, that is,

$$(24)\qquad \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_1) = \max_{i=1,\ldots,K}(\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \hat{\mathrm{L}}_i)).$$

Following this convention and the fact that $e_{l_p}(\sum_{i=2}^K \mathcal{X}_i, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*) = 0$, it is enough to prove that $(\mathrm{L}_1^*, \ldots, \mathrm{L}_K^*)$ is the unique global minimizer w.o.p. of $e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1, \ldots, \mathrm{L}_K)$ in $\mathrm{B}_{\mathrm{G}}((\mathrm{L}_1^*, \ldots, \mathrm{L}_K^*), \gamma_1)$, for sufficiently small γ_1.

Let $t_0 := \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_1)$. For each $1 \leq i \leq K$, we parametrize according to arc length the geodesic lines from L_i^* to $\hat{\mathrm{L}}_i$ by functions $\mathrm{L}_i(t)$, $1 \leq i \leq K$, on the interval $[0, t_0]$ such that

$$(25)\qquad \mathrm{L}_i(0) = \mathrm{L}_i^* \quad \text{and} \quad \mathrm{L}_i(t_0) = \hat{\mathrm{L}}_i.$$

We will prove that for sufficiently small $\gamma_1 > 0$,

$$(26)\quad \frac{\mathrm{d}}{\mathrm{d}t^p}(e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1(t), \ldots, \mathrm{L}_K(t))) > 0 \qquad \text{for all } 0 \leq t \leq \gamma_1 \text{ w.o.p.}$$

This will clearly imply our desired result.

Our proof of (26) is based on the following estimate:

$$(27)\qquad \frac{\mathrm{d}}{\mathrm{d}t^p}(e_{l_p}(\mathbf{x}, \mathrm{L}_1(t), \ldots, \mathrm{L}_K(t)))\Big|_{t=0} \geq -\|\mathbf{x}\|.$$

In order to establish (27), we denote $j = \arg\min_{1\leq i\leq K} \operatorname{dist}(\mathbf{x}, \mathrm{L}_i^*)$ and apply Lemma 2.2 to obtain that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t^p}(e_{l_p}(\mathbf{x}, \mathrm{L}_1(t), \dots, \mathrm{L}_K(t)))\Big|_{t=0} &= \lim_{t\to 0} \frac{\operatorname{dist}(\mathbf{x}, \mathrm{L}_j(t))^p - \operatorname{dist}(\mathbf{x}, \mathrm{L}_j(0))^p}{t^p} \\ &\geq -\|\mathbf{x}\| \lim_{t\to 0} \frac{\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_j(t), \mathrm{L}_j(0))^p}{t^p}.\end{aligned} \tag{28}$$

We also note that for all $0 \leq t \leq t_0$,

$$\frac{\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_j(t), \mathrm{L}_j(0))^p}{t^p} \leq \frac{\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_1(t), \mathrm{L}_1(0))^p}{t^p} = 1. \tag{29}$$

Indeed, if $t = t_0$, the inequality in (29) follows from (24) and the equality follows from (25). Moreover, both of them extend to $0 \leq t < t_0$ by the underlying property of arc length parametrization. Equation (27) thus follows from (28) and (29).

Combining (27) with Hoeffding's inequality, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}t^p}(e_{l_p}(\mathcal{X}_0, \mathrm{L}_1(t), \dots, \mathrm{L}_K(t)))\Big|_{t=0} \geq -\sum_{\mathbf{x}\in\mathcal{X}_0} \|\mathbf{x}\| \geq -\alpha_0 N \qquad \text{w.o.p.} \tag{30}$$

We similarly derive an equation analogous to (30) when replacing $\mathcal{X}_0$ with $\mathcal{X}_1$ by applying some arguments of the proof of Lemma 2.1 and Hoeffding's inequality as follows:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t^p}(e_{l_p}(\mathcal{X}_1, \mathrm{L}_1(t), \dots, \mathrm{L}_K(t)))\Big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}(e_{l_1}(\mathcal{X}_1, \mathrm{L}_1(t)))\Big|_{t=0} \\ &\geq \tau_0\alpha_1 N \qquad \text{w.o.p.}\end{aligned} \tag{31}$$

At last, combining (30), (31) and (6), we obtain that there exists $\gamma_1' \equiv \gamma_1'(D, d, K, p, \alpha_0, \alpha_1)$ such that w.o.p.

$$\frac{\mathrm{d}}{\mathrm{d}t^p}(e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1(t), \dots, \mathrm{L}_K(t)))\Big|_{t=0} \geq (\tau_0\alpha_1 - \alpha_0)N > \gamma_1' N.$$

Using the arguments of the proof of [12], equation (35), we conclude that there exists a constant $\gamma_1 \equiv \gamma_1(D, d, K, p, \alpha_0, \alpha_1, \min_{2\leq i\leq K} \operatorname{dist}(\mathrm{L}_1^*, \mathrm{L}_i^*), \mu_0, \mu_1) > 0$ such that (26) holds.

2.5. *Conclusion of Theorem 1.1.* In order to conclude the theorem, it is enough to prove that $\{\mathrm{L}_1^*, \dots, \mathrm{L}_K^*\}$ is the unique global minimizer w.o.p. of $e_{l_p}(\mathcal{X}_0 \cup \mathcal{X}_1, \mathrm{L}_1, \dots, \mathrm{L}_K)$ in the set

$$\mathrm{GP}(D, d, \gamma_1) := \mathrm{G}(D, d)^K \Big\backslash \bigcup_{i_1, i_2, \dots, i_K \in \mathcal{P}_K} \mathrm{B}_{\mathrm{G}}((\mathrm{L}_{i_1}^*, \dots, \mathrm{L}_{i_K}^*), \gamma_1). \tag{32}$$

Combining Proposition 2.1, the fact that $d_0 > \gamma_1$ [which follows from the definition of d_0 in (13)], Hoeffding's inequality and (6), we obtain that there exists $\gamma_2 \equiv \gamma_2(D, d, K, p, \alpha_0, \min_{1 \le i \le K} \alpha_i, \min_{1 \le i \ne j \le K} \operatorname{dist}(\mathrm{L}_i^*, \mathrm{L}_j^*), \mu_0, \mu_1) > 0$ such that for any fixed $(\hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K) \in \mathrm{GP}(D, d, \gamma_1)$,

$$e_{l_p}(\mathcal{X}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K) - e_{l_p}(\mathcal{X}, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*) > \gamma_2 N \qquad \text{w.o.p.} \tag{33}$$

Following the proof of [12], Theorem 1.1 [i.e., covering $\mathrm{GP}(D, d, \gamma_1)$ by balls], we easily extend (33) w.o.p. for all K subspaces in the set $\mathrm{GP}(D, d, \gamma_1)$ (instead of fixed ones) and thus conclude the theorem.

3. Proof of Theorem 1.2 and a counterexample to asymptotic recovery.

3.1. *Proof of Theorem 1.2.* Following the argument of [12], Section 3.5.1, we reduce the verification of Theorem 1.2 to proving that there exists a constant $\gamma_3 > 0$ such that if for all permutations $i_1, \ldots, i_K \in \mathcal{P}_K$, $\hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K \in \mathrm{G}(D, d)$ satisfy that $\operatorname{dist}_{\mathrm{G}^K}((\mathrm{L}_{i_1}^*, \ldots, \mathrm{L}_{i_K}^*), (\hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K)) > f$, then

$$\mathbb{E}_\mu(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1, \ldots, \hat{\mathrm{L}}_K)) > \mathbb{E}_\mu(e_{l_p}(\mathbf{x}, \mathrm{L}_1^*, \ldots, \mathrm{L}_K^*)) + \gamma_3 + 2\varepsilon^p. \tag{34}$$

In view of Proposition 2.1, in order to conclude (34), it is sufficient to verify that

$$\Bigl(\tau_0 \min_{1 \le j \le K} \alpha_j - \alpha_0\Bigr) f^p > \gamma_3 + 2\varepsilon^p \tag{35}$$

and

$$\tau_0 \min_{1 \le j \le K} \alpha_j \min_{1 \le i, j \le K} \operatorname{dist}_{\mathrm{G}}^p(\mathrm{L}_i^*, \mathrm{L}_j^*)/2^p - \alpha_0 > \gamma_3 + 2\varepsilon^p. \tag{36}$$

Setting $\gamma_3 = \varepsilon^p/2$, (35) follows from (8) and (36) follows from (7).

3.1.1. *Remark on the size of ε.* If

$$\varepsilon > \pi \sqrt{d} 3^{-1/p} \Bigl(\tau_0 \min_{1 \le j \le K} \alpha_j - \alpha_0\Bigr)^{1/p} / 2, \tag{37}$$

then $f > \pi\sqrt{d}/2$, so that there is no restriction on the minimizer of (1) in $\mathrm{G}(D, d)^K$. It thus makes sense to further restrict ε to be at least lower than the right-hand side of (37).

3.2. *A counterexample to exact asymptotic recovery with noise.* One may ask if it is possible in the noisy setting ($\varepsilon > 0$) to recover the underlying subspaces as the number of sampled points, N, approaches infinity. The answer to this question is positive when $K = 1$ (see, e.g., [2], Section 11.6, [18]) or $d = 0$ (see [17]). However, it is often negative when $d > 1$ and $K > 1$, as we demonstrate in Figure 1(a) and explain below. In this example, $D = 2$, $K = 2$, $d = 1$, $\alpha_0 = 0$ and the two underlying distributions μ_1 and μ_2 (corresponding to the two underlying lines L_1^* and L_2^*) are uniformly distributed

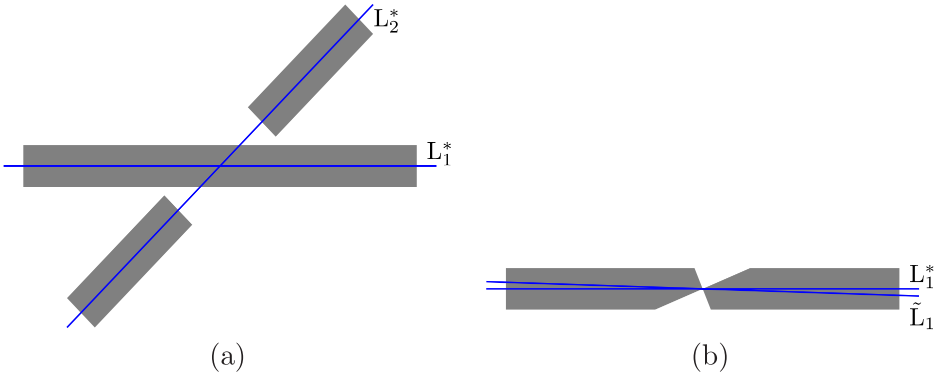

FIG. 1. *A counterexample showing that exact recovery with noise is impossible even asymptotically.* (a) *Gray regions of uniform distributions around the two underlying lines.* (b) *The gray region is the intersection of* Y_1 *with the uniform distribution region around* L_1^*. *The best* l_p *line in* Y_1 *is* $\tilde{L}_1$.

in the two gray regions demonstrated in this figure (the region around L_1^* is a rectangle and the region around L_2^* is a union of two disjoint rectangles).

In order to verify that this is indeed a counterexample, we use a Voronoi-type region, which allows us to reduce approximation by multiple subspaces to approximation by a single subspace on it. Such regions $\{Y_i\}_{i=1}^K$, which are frequently used in Section 4, are obtained by a Voronoi diagram (restricted to the unit ball) of given d-subspaces $\{L_i\}_{i=1}^K \subseteq G(D,d)$ as follows:

$$\begin{aligned}&Y_i(L_1,\dots,L_K)\\ &\quad= \{\mathbf{x} \in B(\mathbf{0},1) : \operatorname{dist}(\mathbf{x}, L_i) < \operatorname{dist}(\mathbf{x}, L_j)\ \forall j : 1 \le j \ne i \le K\}.\end{aligned} \tag{38}$$

These regions are useful to us due to the following elementary proposition, whose trivial proof is described in the Appendix.

PROPOSITION 3.1. *If* $L_1',\dots,L_K' \in G(D,d)$, ν *is a probability measure on* $\mathbb{R}^D$ *and*

$$(L_1',\dots,L_K') = \operatorname*{arg\,min}_{(L_1,\dots,L_K)\in G(D,d)^K} \mathbb{E}_\nu(e_{l_p}(\mathbf{x}, L_1,\dots,L_K)),$$

then

$$L_1' = \operatorname*{arg\,min}_{L_1 \in G(D,d)} \mathbb{E}_\nu(e_{l_p}(\mathbf{x}, L_1) I(\mathbf{x} \in Y_1(L_1', L_2', \dots, L_K'))). \tag{39}$$

We claim that for any fixed $p > 0$, the distance between $\{L_1^*, L_2^*\}$ and the global minimizer of (1) in the setting of this example is bounded from below w.o.p. by a positive constant independent of the sample size, N, for sufficiently large N. Equivalently, we claim that the distance between $\{L_1^*, L_2^*\}$ and the global minimizer of $\mathbb{E}_{\mu_\varepsilon}(\operatorname{dist}^p(\mathbf{x}, \bigcup_{i=1}^K L_i))$ is positive, where μ_ε is the underlying mixture distribution for this example. In view of Proposition 3.1, we only need to show a positive distance between L_1^* and the minimizer of $\mathbb{E}_{\mu_\varepsilon}(e_{l_p}(\mathbf{x},\mathrm{L})I(\mathbf{x}\in \mathrm{Y}_1))$, where $\mathrm{Y}_1 = \mathrm{Y}_1(\mathrm{L}_1^*, \mathrm{L}_2^*)$. We refer to this minimizer as the best l_p line for Y_1 and denote it by $\tilde{\mathrm{L}}_1$ (while arbitrarily fixing p). We note that for any $p > 0$, the integral of l_p distances of points in the part of Y_1 above L_1^* from the line L_1^* is smaller than the similar integral in the bottom part. Therefore, $\tilde{\mathrm{L}}_1$ is different than L_1^* and the respective orientation of the two lines is demonstrated in Figure 1(b). The claim is thus concluded.

4. Proof of Theorem 1.3.

4.1. *Preliminaries.*

4.1.1. *Notation.* We designate the projection from $\mathbb{R}^D$ onto its subspace L by P_{L} and the corresponding orthogonal projection by $P_{\mathrm{L}}^{\perp}$. We define

$$\mathbf{D}_{\mathrm{L},\mathbf{x},p} = P_{\mathrm{L}}(\mathbf{x})P_{\mathrm{L}}^{\perp}(\mathbf{x})^T \operatorname{dist}(\mathbf{x},\mathrm{L})^{(p-2)}. \tag{40}$$

We frequently use the Voronoi-type regions $\{\mathrm{Y}_i\}_{i=1}^K$ defined in (38) with respect to the subspaces $\{\mathrm{L}_i^*\}_{i=1}^K$ and possibly two additional arbitrary subspaces denoted by $\hat{\mathrm{L}}_2 \in \mathrm{G}(D,d)$ and $\tilde{\mathrm{L}}_2 \in \mathrm{G}(D,d)$. We will use the following short notation for $1 \le i \le K$:

$$\hat{\mathrm{Y}}_i = \mathrm{Y}_i(\mathrm{L}_1^*, \hat{\mathrm{L}}_2, \mathrm{L}_3^*, \ldots, \mathrm{L}_K^*), \qquad \tilde{\mathrm{Y}}_i = \mathrm{Y}_i(\mathrm{L}_1^*, \tilde{\mathrm{L}}_2, \mathrm{L}_3^*, \ldots, \mathrm{L}_K^*) \tag{41}$$

and

$$\mathrm{Y}_i = \mathrm{Y}_i(\mathrm{L}_1^*, \mathrm{L}_2^*, \mathrm{L}_3^*, \ldots, \mathrm{L}_K^*). \tag{42}$$

We denote by $\bar{\mathrm{Y}}_i$ the closure of Y_i, that is,

$$\bar{\mathrm{Y}}_i = \{\mathbf{x} \in \mathrm{B}(\mathbf{0},1) : \operatorname{dist}(\mathbf{x},\mathrm{L}_i^*) \le \operatorname{dist}(\mathbf{x},\mathrm{L}_j^*)\ \forall j : 1 \le j \ne i \le K\}. \tag{43}$$

Similarly, the closure of $\hat{\mathrm{Y}}_i$ is denoted by $\bar{\hat{\mathrm{Y}}}_i$.

Let $\mathcal{L}_k$ denote the kth-dimensional Lebesgue measure. We denote $d^* = d \wedge (D-d)$ and let $\theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*)$ be the d^*th largest principal angle between the d-subspaces L_i^* and L_j^*. Our analysis uses the distribution $\mu \equiv \alpha_0\mu_0 + \sum_{i=1}^K \alpha_i\mu_i$, even though the underlying distribution of our model is μ_ε. For L, $\mathrm{L}^* \in \mathrm{G}(D,d)$, we define the "orthogonal subtraction" $\ominus$ as follows:

$$\mathrm{L}^* \ominus \mathrm{L} = \mathrm{L}^* \cap (\mathrm{L} \cap \mathrm{L}^*)^{\perp}.$$

4.1.2. *Auxiliary lemmata.* Using the notation above, we formulate two lemmata, which will be used throughout this proof. The proof of Lemma 4.1 is identical to that of [12], Proposition 2.2 (while replacing sums by expectations), whereas Lemma 4.2 is proved in the Appendix.

LEMMA 4.1. *For any* $\mathrm{L}^* \in \mathrm{G}(D,d)$ *and distribution* μ, *a necessary condition for* L^* *to be a local minimum of* $\mathbb{E}_\mu(l_p(\mathbf{x},\mathrm{L}))$ *is*

$$\mathbb{E}_\mu(\mathbf{D}_{\mathrm{L}^*,\mathbf{x},p}) = \mathbf{0}. \tag{44}$$

The next lemma quantifies the sensitivity of the region Y_j, where $1 \leq j \leq K$, to perturbations in the subspace L_i, where $1 \leq i \neq j \leq K$. WLOG we formulate it with $j = 1$ and $i = 2$ [note that we use the short notation of (41)].

LEMMA 4.2. *If* $\hat{\mathrm{L}}_2, \mathrm{L}_1^*, \mathrm{L}_2^*, \ldots, \mathrm{L}_K^*$ *are subspaces in* $\mathrm{G}(D, d)$ *such that* $\hat{\mathrm{L}}_2 \neq \mathrm{L}_2^*$,

$$\min_{j \neq 2}(\theta_{d^*}(\hat{\mathrm{L}}_2, \mathrm{L}_j^*)) > 0, \qquad \min_{1 \leq i \neq j \leq K}(\theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*)) > 0 \tag{45}$$

and

$$\theta_{d^*}(\hat{\mathrm{L}}_2, \mathrm{L}_1^*) \vee \theta_{d^*}(\mathrm{L}_2^*, \mathrm{L}_1^*) \leq \min_{3 \leq i \leq K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_1^*), \tag{46}$$

then

$$\mathcal{L}_D((\hat{\mathrm{Y}}_1 \setminus \mathrm{Y}_1) \cup (\mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1)) > 0. \tag{47}$$

4.2. *A special case.* The proof of Theorem 1.3 is rather involved. In order to develop a simple intuition, we provide an elementary proof of the very special case where $d = 1$, $p = 2$ and $K = 2$. For simplicity we also assume that $D = 2$, though our argument easily extends to $D > 2$. Figure 2 shows the two underlying lines L_1^* and L_2^* and their corresponding regions Y_1 and Y_2. We note that the best l_2 lines [in $\mathrm{G}(D, 1)$] for μ_0 restricted to Y_1 and Y_2 are the central axes of those regions. Since $\alpha_0 > 0$, the best l_2 lines [in $\mathrm{G}(D, 1)$] for μ restricted to Y_1 and Y_2 (denoted by $\tilde{\mathrm{L}}_1$ and $\tilde{\mathrm{L}}_2$, resp.) must reside between the best l_2 lines for μ_0 restricted to Y_1 and Y_2 and L_1^* and L_2^*, respectively. In particular, they are different from L_1^* and L_2^* as demonstrated in the

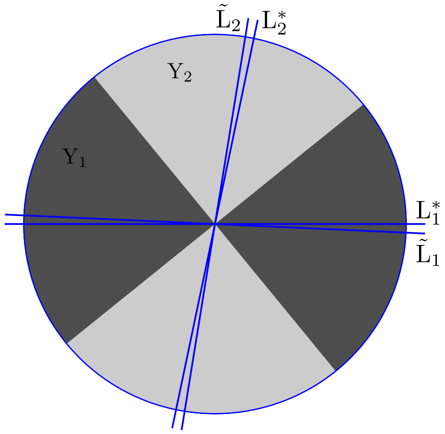

FIG. 2. *Illustrative proof of Theorem 1.3 in the special case where* $p = 2$, $d = 1$, $D = 2$ *and* $K = 2$.

figure. Therefore, $\mathbb{E}_\mu(e_{l_2}(\mathbf{x},\mathrm{L}_1^*,\mathrm{L}_2^*)) > \mathbb{E}_\mu(e_{l_2}(\mathbf{x},\tilde{\mathrm{L}}_1,\tilde{\mathrm{L}}_2))$. This implies that w.o.p. $e_{l_2}(\mathcal{X},\mathrm{L}_1^*,\mathrm{L}_2^*) > e_{l_2}(\mathcal{X},\tilde{\mathrm{L}}_1,\tilde{\mathrm{L}}_2)$.

4.3. *Reduction of the statement of Theorem 1.3 to simpler formulations.*

4.3.1. *Reduction* I*: Using the Voronoi-type regions* $\{\mathrm{Y}_i\}_{i=1}^K$. We will show here that the following equation implies Theorem 1.3:

$$\gamma_{D,d}^K(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : \mathbb{E}_{\mu_0}(I(\mathbf{x}\in \mathrm{Y}_j)\mathbf{D}_{\mathrm{L}_j^*,\mathbf{x},p}) = \mathbf{0}\ \forall 1\le j\le K) = 0. \tag{48}$$

First, we apply the argument of [12], Section 3.6.1 (which requires the assumption specified in Section 1.3 that the first moments of $\{\|\nu_{i,\varepsilon}\|\}_{i=1}^K$ are smaller than ε) to obtain that Theorem 1.3 follows by the equation

$$\begin{aligned}\gamma_{D,d}^K\Big(&\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : (\mathrm{L}_1^*,\dots,\mathrm{L}_K^*)\\ &= \operatorname*{arg\,min}_{(\mathrm{L}_1,\dots,\mathrm{L}_K)} \mathbb{E}_\mu(e_{l_p}(\mathbf{x},\mathrm{L}_1,\dots,\mathrm{L}_K))\Big) = 0.\end{aligned} \tag{49}$$

Next, applying Proposition 3.1, we conclude that (49) is a direct consequence of the equation:

$$\begin{aligned}\gamma_{D,d}^K\Big(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : \mathrm{L}_j^* = \operatorname*{arg\,min}_{\mathrm{L}\in\mathrm{G}(D,d)} \mathbb{E}_\mu(e_{l_p}(\mathbf{x},\mathrm{L})I(\mathbf{x}\in\mathrm{Y}_j))&\\ \forall 1\le j\le K\Big) = 0.&\end{aligned} \tag{50}$$

Furthermore, applying Lemma 4.1 with $\mu = \mu|_{\mathrm{Y}_j}$, we obtain that (50) follows by the equation

$$\gamma_{D,d}^K(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : \mathbb{E}_{\mu}(I(\mathbf{x}\in \mathrm{Y}_j)\mathbf{D}_{\mathrm{L}_j^*,\mathbf{x},p}) = 0\ \forall 1\le j\le K) = 0. \tag{51}$$

At last we conclude the desired reduction by noting that (51) and (48) are equivalent [indeed, the only relevant components of the distribution μ in (51) are μ_0 and μ_j and the corresponding expectation according to μ_j is zero].

4.3.2. *Reduction* II*: From K subspaces to a single subspace.* We reduce (48) so that its underlying condition involves a single subspace as follows:

$$\begin{aligned}\gamma_{D,d}\Big(&\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_j^*) > 0,\\ &\operatorname*{arg\,min}_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*) = 2, \mathbb{E}_{\mu_0}(I(\mathbf{x}\in\mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0}\Big) = 0.\end{aligned} \tag{52}$$

We remark that some of the underlying technical conditions of (52) appear in (45) and (46) and will be better understood later when applying Lemma 4.2.

We verify this reduction as follows. WLOG (52) can be formulated by replacing L_2^* with L_k^*, for some $3\le k\le K$, while letting $\operatorname{arg\,min}_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*,$

$\mathrm{L}_i^*) = k$. Combining this observation with elementary properties of distributions, we have that

$$\begin{aligned}
&\gamma_{D,d}^K(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_j)\mathbf{D}_{\mathrm{L}_j^*,\mathbf{x},p}) = \mathbf{0} \ \forall 1 \le j \le K) \\
&\quad \le \sum_{k=2}^K \int_{\mathrm{G}(D,d)^{K-1}} \gamma_{D,d}\Big(\mathrm{L}_k^* : \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0, \\
&\qquad\qquad \operatorname*{arg\,min}_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = k, \\
&\qquad\qquad \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0} | \{\mathrm{L}_i^*\}_{1\le i\ne k\le K}\Big) \mathrm{d}(\gamma_{D,d}^{K-1}(\{\mathrm{L}_i^*\}_{1\le i\ne k\le K})) \\
&\qquad + \gamma_{D,d}^K\Big(\{\mathrm{L}_i^*\}_{i=1}^K \subset \mathrm{G}(D,d) : \min_{1\le i,j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) = 0\Big) = 0.
\end{aligned}$$

4.4. *Concluding the cases $d = 1$ and $d = D - 1$.* We assume first that $d = 1$. We conclude the theorem in this case by proving (52) and then extend the analysis to the case $d = D - 1$.

4.4.1. *Reduction of (52) using additional condition on the Grassmannian.* We fix $\mathbf{v}_1$ to be one of the two unit vectors spanning L_1^* and denote by $\mathbf{u}_1$ the unit vector spanning $(\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp}$ having orientation such that for any point $\mathbf{x} \in \mathrm{L}_2^* : (\mathbf{x}^T\mathbf{u}_1)(\mathbf{x}^T\mathbf{v}_1) \ge 0$. We will prove that (52) follows from the following equation, which introduces a restriction on the Grassmannian:

$$\begin{aligned}
&\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0, \\
&\qquad \operatorname*{arg\,min}_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\
&\qquad \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0} | (\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp} = \mathrm{Sp}(\mathbf{u}_1)\Big) = 0.
\end{aligned} \tag{53}$$

We define the following subset of the sphere $S^{D-1} : \Omega_0 = \{\mathbf{x} \in S^{D-1} : \mathbf{x} \perp \mathbf{v}\}$, and a distribution ω on Ω_0 such that for any $\mathrm{A} \subseteq \Omega_0 : \omega(\mathrm{A}) = \gamma_{D,d}(\mathrm{L}_2^* \in \mathrm{G}(D,d) : (\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp} \in \mathrm{Sp}(\mathrm{A}))$. Using this notation, (53) implies (52) as follows:

$$\begin{aligned}
&\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0, \operatorname*{arg\,min}_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\
&\qquad\qquad\qquad\qquad \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0}\Big) \\
&\quad = \int_{\Omega_0} \gamma_{D,d}\Big(\mathrm{L}_2^* : \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0, \operatorname*{arg\,min}_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\
&\qquad\qquad \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0} | (\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^{*\perp} = \mathrm{Sp}(\mathbf{u}_1)\Big) \mathrm{d}(\omega(\mathbf{u}_1)) \\
&\quad = 0.
\end{aligned}$$

4.4.2. *Proof of (53).* We will show that at most one element satisfies the underlying condition of (53) (i.e., it is a member of the set for which $\gamma_{D,d}$ is evaluated). Assume, on the contrary, that there are two subspaces $\hat{\mathrm{L}}_2$ and $\tilde{\mathrm{L}}_2$ satisfying this condition with corresponding angles $\hat{\theta} = \theta_{d^*}(\mathrm{L}_1^*, \hat{\mathrm{L}}_2)$ and $\tilde{\theta} = \theta_{d^*}(\mathrm{L}_1^*, \tilde{\mathrm{L}}_2)$ in $[0, \pi/2]$, where WLOG $\hat{\theta} > \tilde{\theta}$. Using the notation of (41), we have that

$$\begin{aligned} &\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) - \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) \\ &\quad = 2 \cdot (\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \tilde{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) - \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \hat{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p})) \\ &\quad = \mathbf{0} - \mathbf{0} = \mathbf{0}. \end{aligned} \tag{54}$$

Consequently,

$$\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1)\mathbf{v}_1^T\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\mathbf{u}_1) - \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1)\mathbf{v}_1^T\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}\mathbf{u}_1) = \mathbf{0}. \tag{55}$$

Defining

$$\theta_{\mathbf{u}_1,\mathbf{v}_1}(\mathbf{x}) = \arctan\frac{\mathbf{u}_1 \cdot \mathbf{x}}{\mathbf{v}_1 \cdot \mathbf{x}}$$

and

$$\mathrm{Y}_{1,\hat{2}} = \left\{\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) : \mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \min_{3 \le i \le K} \mathrm{dist}(\mathbf{x}, \mathrm{L}_i^*)\right\},$$

we express the regions $\hat{\mathrm{Y}}_1$ and $\tilde{\mathrm{Y}}_1$ as follows:

$$\hat{\mathrm{Y}}_1 = \mathrm{Y}_{1,\hat{2}} \cap \{\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) : \hat{\theta}/2 - \pi/2 < \theta_{\mathbf{u}_1,\mathbf{v}_1}(\mathbf{x}) < \hat{\theta}/2\}, \tag{56}$$

$$\tilde{\mathrm{Y}}_1 = \mathrm{Y}_{1,\hat{2}} \cap \{\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) : \tilde{\theta}/2 - \pi/2 < \theta_{\mathbf{u}_1,\mathbf{v}_1}(\mathbf{x}) < \tilde{\theta}/2\}. \tag{57}$$

Figure 3 clarifies (56) and (57) in the special case where $d = 1$ and $K = 2$.

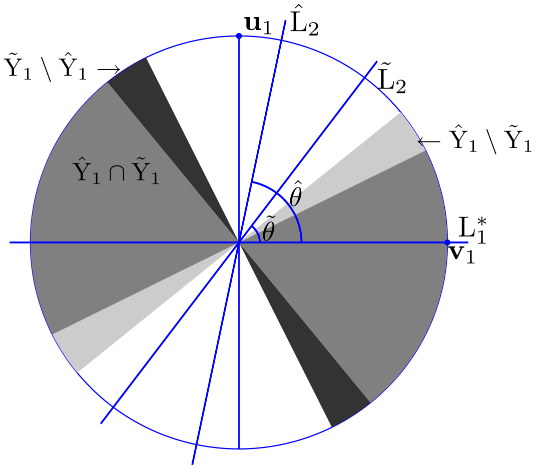
$\mathbf{u}_1$
$\hat{\mathrm{L}}_2$
$\tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1$ →
$\tilde{\mathrm{L}}_2$
← $\hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1$
$\hat{\mathrm{Y}}_1 \cap \tilde{\mathrm{Y}}_1$
$\hat{\theta}$
$\tilde{\theta}$
L_1^*
$\mathbf{v}_1$

FIG. 3. *The regions $\hat{\mathrm{Y}}_1$ and $\tilde{\mathrm{Y}}_1$ and the relation to $\hat{\theta}$ and $\tilde{\theta}$ when $d = 1$ and $K = 2$.*

Combining (56) and (57) with the definition of $\mathbf{D}_{\mathrm{L},\mathbf{x},p}$ in (40), we obtain that

$$\hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1 \subset \{\mathbf{x} \in \mathrm{B}(\mathbf{0},1) : \mathbf{v}_1^T \mathbf{x}\mathbf{x}^T \mathbf{u}_1 \equiv \operatorname{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(2-p)} \mathbf{v}_1^T \mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p} \mathbf{u}_1 > 0\} \tag{58}$$

and

$$\tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1 \subset \{\mathbf{x} \in \mathrm{B}(\mathbf{0},1) : \mathbf{v}_1^T \mathbf{x}\mathbf{x}^T \mathbf{u}_1 \equiv \operatorname{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(2-p)} \mathbf{v}_1^T \mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p} \mathbf{u}_1 < 0\}. \tag{59}$$

It follows from Lemma 4.2 that $\mathcal{L}_D((\tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1) \cup (\hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1)) > 0$ and, consequently, for any $r > 0$, $\mathcal{L}_D(\mathrm{B}(\mathbf{0},r) \cap ((\tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1) \cup (\hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1))) > 0$ (indeed, if $\mathbf{x} \in \mathrm{Y}_1$, then $c \cdot \mathbf{x} \in \mathrm{Y}_1$ for any $0 < c < 1/\|\mathbf{x}\|$; thus, the distribution in the latter inequality is just a scaling by r^D of the distribution in the former one). Since there exists $r > 0$ such that the restriction of $\mathcal{L}_D$ to $\mathrm{B}(\mathbf{0},r)$ is absolutely continuous with respect to μ_0, we also have that $\mu_0(\mathrm{B}(\mathbf{0},r) \cap ((\tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1) \cup (\hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1))) > 0$. However, this contradicts (55), (58) and (59), that is, it proves (53) and therefore the theorem in the current special case.

4.4.3. *The case $d = D - 1$.* We note that the proof of the above case ($d = 1$) can be adapted to the case where $d = D - 1$. This is done by letting $\mathbf{v}_1$ be one of the two unit vectors spanning $\mathrm{L}_1^* \cap (\mathrm{L}_1^* \cap \mathrm{L}_2^*)^\perp$ [note that $\dim(\mathrm{L}_1^*) = D - 1$ and $\dim(\mathrm{L}_1^* \cap \mathrm{L}_2^*) = d - 2$ so that $\dim(\mathrm{L}_1^* \cap (\mathrm{L}_1^* \cap \mathrm{L}_2^*)^\perp) = 1$] and $\mathbf{u}_1$ be the unit vector of $(\mathrm{L}_1^* + \mathrm{L}_2^*) \cap \mathrm{L}_1^\perp$ with a similar orientation as in the case where $d = 1$.

4.5. *Conclusion: The case where $d \neq 1$ and $d \neq D - 1$.*

4.5.1. *Reduction of (52) using additional condition on the Grassmannian.* The following reduction is analogous to the one of Section 4.4.1. Denoting by $B(\mathbb{R}^D, \mathbb{R}^D)$ the space of linear operators from $\mathbb{R}^D$ to itself, we define

$$\Omega_1 = \{(P_1, P_2) \in B(\mathbb{R}^D, \mathbb{R}^D)^2 : \exists \mathrm{L} \in \mathrm{G}(D,d) \text{ not orthogonal to } \mathrm{L}_1^*, \\ \text{s.t. } \dim(\mathrm{L}_1^* \ominus \mathrm{L}) > 1, P_{\mathrm{L}_1^*}^T P_{\mathrm{L}} P_{\mathrm{L}_1^*} = P_1, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}} P_{\mathrm{L}_1^*}^\perp = P_2\}$$

and the distribution ω_1 on Ω_1 as follows: for any set $\mathrm{A} \subseteq \Omega_1$,

$$\omega_1(\mathrm{A}) = \gamma_{D,d}(\mathrm{L} \in \mathrm{G}(D,d) : (P_{\mathrm{L}_1^*}^T P_{\mathrm{L}} P_{\mathrm{L}_1^*}, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}} P_{\mathrm{L}_1^*}^\perp) \in \mathrm{A}).$$

Using this notation, we reduce (52) as follows:

$$\begin{aligned} \gamma_{D,d}\Big(&\mathrm{L}_2^* \in \mathrm{G}(D,d) : \mathrm{L}_1^* \not\perp \mathrm{L}_2^*, \dim(\mathrm{L}_1^* \cap \mathrm{L}_2^{*\perp}) > 1, \\ &\min_{1 \le i \ne j \le K} \theta_{d^*}(\mathrm{L}_i^*, \mathrm{L}_j^*) > 0, \operatorname*{arg\,min}_{2 \le i \le K} \theta_{d^*}(\mathrm{L}_1^*, \mathrm{L}_i^*) = 2, \\ &\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \mathrm{Y}_1) \mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0} | \\ &(P_{\mathrm{L}_1^*}^T P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}^\perp) = (P_1, P_2) \in \Omega_1\Big) = 0. \end{aligned} \tag{60}$$

Indeed,

$$
\begin{aligned}
&\gamma_{D,d}\Big(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_j^*) > 0, \arg\min_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*) = 2,\\
&\qquad\qquad\qquad\qquad\qquad\qquad \mathbb{E}_{\mu_0}(I(\mathbf{x}\in \mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0}\Big)\\
&\quad \le \int_{\Omega_1} \gamma_{D,d}(\mathrm{L}_2^* : \mathrm{L}_1^* \text{ is not orthogonal to } \mathrm{L}_2^*,\\
&\qquad\qquad\qquad \dim(\mathrm{L}_1^* \ominus \mathrm{L}_2^*) > 1, \min_{1\le i\ne j\le K} \theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_j^*) > 0,\\
&\qquad\qquad\qquad \arg\min_{2\le i\le K} \theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*) = 2,\\
&\qquad\qquad\qquad \mathbb{E}_{\mu_0}(I(\mathbf{x}\in \mathrm{Y}_1)\mathbf{D}_{\mathrm{L}_1^*,\mathbf{x},p}) = \mathbf{0}|\\
&\qquad\qquad\qquad\quad (P_{\mathrm{L}_1^*}^T P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}, P_{\mathrm{L}_1^*}^{\perp T} P_{\mathrm{L}_2^*} P_{\mathrm{L}_1^*}^{\perp}) = (P_1,P_2) \in \Omega_1)\,\mathrm{d}(\omega_1(P_1,P_2))\\
&\qquad + \gamma_{D,d}(\mathrm{L}_2^* \in \mathrm{G}(D,d) : \dim(\mathrm{L}_1^* \ominus \mathrm{L}_2^*) \le 1, \text{ or } \mathrm{L}_2^* \perp \mathrm{L}_1^*) = 0 + 0 = 0.
\end{aligned}
$$

4.5.2. *Bulk of the proof.* We prove (60) by using the following two lemmata, which are proved below (Sections 4.5.3 and 4.5.4).

LEMMA 4.3. *If* $\dim(\mathrm{L}_1^* \ominus \mathrm{L}_2^*) \ge 2$ *and* L_1^* *is not orthogonal to* L_2^*, *then the set*

$$\mathrm{Z} = \{\mathrm{L} \in \mathrm{G}(D,d) : P_{\mathrm{L}_1^*}(P_{\mathrm{L}_2^*} - P_{\mathrm{L}})P_{\mathrm{L}_1^*} = 0, P_{\mathrm{L}_1^*}^{\perp}(P_{\mathrm{L}_2^*} - P_{\mathrm{L}})P_{\mathrm{L}_1^*}^{\perp} = 0\}$$

is infinite.

LEMMA 4.4. *If* $\tilde{\mathrm{L}}_2, \hat{\mathrm{L}}_2 \in \mathrm{G}(D,d)$ *satisfy* $\tilde{\mathrm{L}}_2 \ne \hat{\mathrm{L}}_2$, $\theta_{d^*}(\hat{\mathrm{L}}_2,\mathrm{L}_1^*) \vee \theta_{d^*}(\mathrm{L}_2^*,\mathrm{L}_1^*) \le \min_{3\le i\le K}\theta_{d^*}(\mathrm{L}_i^*,\mathrm{L}_1^*)$, $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*} = 0$ *and* $P_{\mathrm{L}_1^*}^{\perp}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}^{\perp} = 0$, *then either* $\hat{\mathrm{L}}_2$ *or* $\tilde{\mathrm{L}}_2$ *will not satisfy the condition in (60).*

To conclude (60), we rewrite it as follows: $\gamma_{D,d}(A|B) = 0$, where A and B are clear from the context. We note that Lemma 4.3 implies that there are infinitely many subspaces L_2^* in B. On the other hand, Lemma 4.4 implies that there is only one subspace L_2^* in A. These observations clearly prove (60). We remark that the idea of this proof is somewhat similar to that of the previous case where $d = 1$ or $d = D - 1$. In this case, Lemma 4.3 is analogous to the fact that there is a degree of freedom in choosing L_2^* in (53) [since we can choose any $\theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_2^*) < \min_{3\le i\le K}\theta_{d^*}(\mathrm{L}_1^*,\mathrm{L}_i^*)$]. Moreover, Lemma 4.4 is analogous to the fact that there were not two subspaces $\hat{\mathrm{L}}_2$ and $\tilde{\mathrm{L}}_2$ satisfying the underlying condition of (53).

4.5.3. *Proof of Lemma 4.3.* We denote $\tilde{\mathrm{L}}_1 = \mathrm{L}_1^* \ominus (\mathrm{L}_1^* \cap \mathrm{L}_2^*)$ and $\tilde{\mathrm{L}}_2 = \mathrm{L}_2^* \ominus (\mathrm{L}_1^* \cap \mathrm{L}_2^*)$. The idea of the proof is to construct a one-to-one function $g : S^{D-1} \cap \tilde{\mathrm{L}}_2 \to \mathrm{Z}$. Then, using this function and the fact that $\dim(\tilde{\mathrm{L}}_2) = \dim(\mathrm{L}_1^*) - \dim(\mathrm{L}_2^* \cap \mathrm{L}_1^*) \ge 2$, we conclude that Z, which contains $g(S^{D-1} \cap \tilde{\mathrm{L}}_2)$, is infinite.

For any $\mathbf{u}_0 \in S^{D-1} \cap \tilde{\mathrm{L}}_2$, we arbitrarily fix $\mathbf{v}_0 = \mathbf{v}_0(\mathbf{u}_0)$ as one of the two unit vectors spanning $\tilde{\mathrm{L}}_1 \cap (\tilde{\mathrm{L}}_2 \ominus \mathrm{Sp}(\mathbf{u}_0))^{\perp}$. The vector $\mathbf{v}_0$ exists since

$$\begin{aligned}\dim(\tilde{\mathrm{L}}_1 \cap (\tilde{\mathrm{L}}_2 \ominus \mathrm{Sp}(\mathbf{u}_0))^{\perp}) &\geq \dim(\tilde{\mathrm{L}}_1) + \dim((\tilde{\mathrm{L}}_2 \ominus \mathrm{Sp}(\mathbf{u}_0))^{\perp}) - D \\ &= d + (D - d + 1) - D = 1.\end{aligned}$$

We define the function g as follows:

$$g(\mathbf{u}_0) = \mathrm{Sp}(\mathbf{u}_0 - 2(\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0, \mathrm{L}_2^* \ominus \mathrm{Sp}(\mathbf{u}_0)).$$

We first claim that the image of g is contained in Z. Indeed, we note that

$$\begin{aligned}P_{g(\mathbf{u}_0)} - P_{\mathrm{L}_2^*} &= (\mathbf{u}_0 - 2(\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0)^T (\mathbf{u}_0 - 2(\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0) - \mathbf{u}_0^T \mathbf{u}_0 \\ &= -2(\mathbf{v}_0^T \mathbf{u}_0)(\mathbf{v}_0^T(\mathbf{u}_0 - (\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0) + (\mathbf{u}_0 - (\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0)^T . \mathbf{v}_0)\end{aligned} \tag{61}$$

Combining (61) with the following two facts: $\mathbf{v}_0 \in \mathrm{L}_1^*$ and $\mathbf{u}_0 - (\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0 \in \mathrm{L}_1^{*\perp}$, we obtain that $g(\mathbf{u}_0) \in \mathrm{Z}$.

At last, we prove that g is one-to-one and thus conclude the proof. If, on the contrary, there exist $\mathbf{u}_1$, $\mathbf{u}_2 \in S^{D-1} \cap \tilde{\mathrm{L}}_2$ such that $\mathbf{u}_1 \neq \mathbf{u}_2$ and $g(\mathbf{u}_1) = g(\mathbf{u}_2)$, then $g(\mathbf{u}_1) = \mathrm{Sp}(g(\mathbf{u}_1), g(\mathbf{u}_2)) \supseteq (\mathrm{L}_2^* \ominus \mathrm{Sp}(\mathbf{u}_1)) + (\mathrm{L}_2^* \ominus \mathrm{Sp}(\mathbf{u}_2)) \supseteq \mathrm{L}_2^*$. Since $\dim(g(\mathbf{u}_1)) = \dim(\mathrm{L}_2^*)$, we conclude that $g(\mathbf{u}_1) = \mathrm{L}_2^*$. On the other hand, we claim that for any $\mathbf{u}_0 \in S^{D-1} \cap \tilde{\mathrm{L}}_2 : g(\mathbf{u}_0) \neq \mathrm{L}_2^*$ and thus obtain a contradiction. Indeed, since $\mathbf{u}_0 \in \tilde{\mathrm{L}}_2$, $\mathbf{v}_0 \in \tilde{\mathrm{L}}_1$ and L_1^* is not orthogonal to L_2^*, we have that $\mathbf{v}_0^T \mathbf{u}_0 \neq 0$ and, consequently, $\mathbf{u}_0 - (\mathbf{v}_0^T \mathbf{u}_0)\mathbf{v}_0 \neq \mathbf{u}_0$. Applying the latter observation in (61), we obtain that $P_{g(\mathbf{u}_0)} \neq P_{\mathrm{L}_2^*}$ and, consequently, $g(\mathbf{u}_0) \neq \mathrm{L}_2^*$.

4.5.4. *Proof of Lemma 4.4.* We assume, on the contrary, that both $\hat{\mathrm{L}}_2$ and $\tilde{\mathrm{L}}_2$ satisfy the underlying condition of (52) and conclude a contradiction.

We arbitrarily fix here $\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1$ [using the notation of (41)]. We note that $\mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_2)$ and $\mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \arg\min_{3 \leq i \leq K} \mathrm{dist}(\mathbf{x}, \mathrm{L}_i^*)$. Since $\mathbf{x} \notin \tilde{\mathrm{Y}}_1$, we have that $\mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) > \mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}_2)$ and, thus,

$$\mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}_2) < \mathrm{dist}(\mathbf{x}, \mathrm{L}_1^*) < \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_2). \tag{62}$$

Consequently,

$$\mathbf{x}^T (P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})\mathbf{x} = \mathrm{dist}(\mathbf{x}, \tilde{\mathrm{L}}_2)^2 - \mathrm{dist}(\mathbf{x}, \hat{\mathrm{L}}_2)^2 < 0. \tag{63}$$

We partition $P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2}$ into four parts: $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}$, $P_{\mathrm{L}_1^*}^{\perp}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}^{\perp}$, $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}^{\perp}$ and $P_{\mathrm{L}_1^*}^{\perp}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}$. The first two are zero, and the last two are adjoint to each other; we thus only consider $P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}^{\perp}$. Let its SVD be

$$P_{\mathrm{L}_1^*}(P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2})P_{\mathrm{L}_1^*}^{\perp} = \mathbf{U\Sigma V} = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{64}$$

We can express the SVD of $P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2}$ using (64) and the partition above as follows:

$$P_{\hat{\mathrm{L}}_2} - P_{\tilde{\mathrm{L}}_2} = \sum_{i=1}^{d} \sigma_i (\mathbf{u}_i \mathbf{v}_i^T + \mathbf{v}_i \mathbf{u}_i^T). \tag{65}$$

Combining (63) and (65), we obtain that

$$\sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T \mathbf{x}\mathbf{x}^T \mathbf{v}_i = \mathbf{x}^T \left(\sum_{i=1}^{n} \sigma_i (\mathbf{u}_i \mathbf{v}_i^T + \mathbf{v}_i \mathbf{u}_i^T) \right) \mathbf{x}/2 < 0. \tag{66}$$

We define a function $f : \mathbb{R}^{D \times D} \to \mathbb{R}$ such that for any $\mathbf{A} \in \mathbb{R}^{D \times D} : f(\mathbf{A}) = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T \mathbf{A} \mathbf{v}_i$. Using (66) and the fact that $\{\mathbf{u}_i\}_{i=1}^{d} \in \mathrm{L}_1^*$ and $\{\mathbf{v}_i\}_{i=1}^{d} \in \mathrm{L}_1^{*\perp}$, we deduce that

$$\begin{aligned} f(\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) &= \operatorname{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(p-2)} f(P_{\mathrm{L}_1^*}(\mathbf{x}) P_{\mathrm{L}_1^*}^{\perp}(\mathbf{x})^T) \\ &= \operatorname{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(p-2)} \sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T P_{\mathrm{L}_1^*}(\mathbf{x}) P_{\mathrm{L}_1^*}^{\perp}(\mathbf{x})^T \mathbf{v}_i \\ &= \operatorname{dist}(\mathbf{x}, \mathrm{L}_1^*)^{(p-2)} \sum_{i=1}^{n} \sigma_i \mathbf{u}_i^T \mathbf{x}\mathbf{x}^T \mathbf{v}_i < 0. \end{aligned} \tag{67}$$

Similarly, for any point $\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1$,

$$f(\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) > 0. \tag{68}$$

Combining (54), (67), (68), Lemma 4.2 and the linearity of f, we conclude the following contradiction establishing the current lemma:

$$\begin{aligned} 0 &= f(\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p}) - \mathbb{E}_{\mu_0}(I(\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p})) \\ &= f(\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \tilde{\mathrm{Y}}_1 \setminus \hat{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p})) - f(\mathbb{E}_{\mu_0}(I(\mathbf{x} \in \hat{\mathrm{Y}}_1 \setminus \tilde{\mathrm{Y}}_1)\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p})) \\ &> 0. \end{aligned} \tag{69}$$

4.5.5. *Remark on the sizes of* δ_0 *and* κ_0. The constants δ_0 and κ_0 depend on other parameters of the underlying weak HLM model, in particular, the underlying subspaces $\{\mathrm{L}_i^*\}_{i=1}^{K}$. For example, one can bound both κ_0 and δ_0 from below by the following number:

$$\max_{1 \le i \le K} \Big(\mathbb{E}_\mu(e_{l_p}(\mathbf{x}, \mathrm{L}_i^*) I(\mathbf{x} \in \mathrm{Y}_i)) - \min_{\mathrm{L} \in \mathrm{G}(D,d)} \mathbb{E}_\mu(e_{l_p}(\mathbf{x}, \mathrm{L}) I(\mathbf{x} \in \mathrm{Y}_i)) \Big) / (4p).$$

If $p \ge 2$, then a simpler lower bound on both κ_0 and δ_0 is

$$\frac{\| \max_{1 \le i \le K} \mathbb{E}_\mu(\mathbf{D}_{\mathrm{L}_1^*, \mathbf{x}, p} I(\mathbf{x} \in \mathrm{Y}_i)) \|_2^2}{pdD2^{p+5}}.$$

5. Discussion.

We studied the effectiveness of l_p minimization for recovering (or nearly recovering) all underlying K subspaces for i.i.d. samples from two different types of HLM distributions. In particular, we demonstrated a phase transition phenomenon around $p=1$.

We discuss here implications, extensions and limitations of this theory as well as some open directions.

5.1. *Obstacles for convex recovery of multiple subspaces.*

There are some recent methods for robust single subspace recovery by convex optimization (see, e.g., [6]). Such methods minimize a real-valued convex function h on a convex set $\mathbb{H}$ (e.g., set of matrices), which can be mapped on $\mathrm{G}(D,d)$. However, such a minimization cannot be done for multiple subspaces. Indeed, in that case one must minimize a multivariate function $h:\mathbb{H}^K \to \mathbb{R}$ for convex $\mathbb{H}$. Clearly, the function h must be invariant to permutations of coordinates. Let g be a mapping of $\mathbb{H}$ onto $\mathrm{G}(D,d)$. It follows from the assumption that the minimization of h leads to the underlying subspaces $\{\mathrm{L}_i^*\}_{i=1}^K$ and the permutation-invariance of h that the set of minimizers of h coincides with all permutations of $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_K$, where $\hat{\mathbf{x}}_i \in g^{-1}(\mathrm{L}_i^*)$ for all $1 \le i \le K$. Since h is convex, $(\sum_{i=1}^K \hat{\mathbf{x}}_i/K, \ldots, \sum_{i=1}^K \hat{\mathbf{x}}_i/K)$ is also a minimizer of h. Consequently, $\sum_{i=1}^K \hat{\mathbf{x}}_i/K \in g^{-1}(\mathrm{L}_j^*)$ for all $1 \le j \le K$, and, thus, $g(\sum_{i=1}^K \hat{\mathbf{x}}_i/K) = \mathrm{L}_1^* = \cdots = \mathrm{L}_K^*$, which is a contradiction.

Furthermore, a minimization on $\mathrm{G}(D,d)^K$ cannot even be geodesically convex. Indeed, the maximum of a geodesically convex function on a compact, geodesically convex set is attained on the boundary. However, $\mathrm{G}(D,d)^K$ is compact, geodesically convex and has no boundary, so any function defined on $\mathrm{G}(D,d)^K$ is not geodesically convex.

5.2. *Implications for a single subspace recovery.*

In [12], we discussed the recovery of a single subspace. Theorems 1.1 and 1.2 apply to this case when $K=1$. Unlike [12] which assumed that μ_0 was spherically symmetric (while having possibly additional "outliers" along other subspaces, distributed according to $\{\mu_i\}_{i=2}^K$), here we have a very weak requirement from μ_0 (which represents all outliers). However, here there is a strong restriction on the fraction of outliers, α_0, whereas in [12] there was no requirement, except for $\alpha_0 < 1$.

5.3. *Extending our theory to more general distributions.*

In Theorems 1.1 and 1.2, the strict spherical symmetry of $\{\mu_i\}_{i=1}^K$ (within $\{\mathrm{L}_i\}_{i=1}^K$, resp.) can be replaced by approximate spherical symmetry of $\{\mu_i\}_{i=1}^K$. That is, for each $1 \le i \le K$ and L_i and μ_i as before, we form a new distribution μ_i', with the same support as μ_i such that the derivative of μ_i' w.r.t. μ_i is bounded away from 0 and ∞. We then replace μ_i with μ_i'. This new setting will require replacing $\{\alpha_i\}_{i=1}^K$ in (6)–(8) by $\{\delta_i\,\alpha_i\}_{i=1}^K$, where $\delta_i \equiv \delta_i(\mu_i', \mu_i)$ for $1 \leq i \leq K$ (δ_i is the lowest value of the derivative of μ_i' w.r.t. μ_i).

Furthermore, the boundedness of the support of the distributions $\{\mu_i\}_{i=0}^K$ can be weakened by assuming that these distributions are sub-Gaussian. Indeed, this will mainly require changing Hoeffding's inequality with [19], Proposition 2.1.9.

5.4. *Distributions resulting in counterexamples for our theory.* There are several typical cases with settings different than above, where the underlying subspaces cannot be recovered by minimizing the energy (1) for all $p > 0$.

The first typical example is when there is an outlier with sufficiently large magnitude so that the minimizer of (1) contains a subspace passing through this outlier, which is different than any of the underlying subspaces. Our setting avoids such a counterexample by requiring (6). We briefly provide the idea as follows: an arbitrarily large outlier in our setting of supports within $\mathrm{B}(\mathbf{0}, 1)$ means, for example, that the outlier has magnitude one and the inliers are supported within $\mathrm{B}(\mathbf{0}, \varepsilon)$, where ε is arbitrarily small. Therefore, $\psi(\varepsilon) = 1$, so that $\psi_{\mu_1}^{-1}((1 + (2K-1)\mu_1(\{\mathbf{0}\}))/2K) < \psi_{\mu_1}^{-1}(1) = \varepsilon$ and, consequently, $\tau_0 \lessapprox \varepsilon^p$. In view of (6), we control the fraction of outliers as a function of ε^p. In particular, for a fixed sample size and sufficiently small ε, no outliers are allowed by this condition.

The second example is when the distribution of outliers lies on another subspace, $\mathrm{L}_0^* \in \mathrm{G}(D, d)$ and $\alpha_0 > \min_{1 \leq i \leq K} \alpha_i$, so that L_0^* is contained in the minimizer of (1). Our setting avoids this counterexample by assuming an upper bound on the percentage of outliers in terms of the minimal percentage of inliers [see (6)].

For the last example we assume for simplicity that $D = 2$, $d = 1$, $K = 2$ and underlying uniform distributions (of outliers and along the two underlying lines) restricted to the unit disk. We further assume that the two lines have angles ε and $-\varepsilon$ w.r.t. the x-axis. By choosing ε sufficiently small the x-axis and y-axis provide a smaller value for the energy (1) than the underlying lines. We note that in this case (6) does not hold [due to the small size of $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_i^*, \mathrm{L}_j^*)$].

5.5. *Another phase transition at $p = 1$: Many local minima for $0 < p < 1$.* Our previous work [12], proof of Proposition 2.1, implies that if $0 < p < 1$ and there exist distinct subspaces $\{\mathrm{L}_i\}_{i=1}^K \subseteq \mathrm{G}(D, d)$ such that $\mathrm{Sp}(\mathcal{X} \cap \mathrm{L}_i) = \mathrm{L}_i$ for all $1 \leq i \leq K$, then $\{\mathrm{L}_i\}_{i=1}^K$ is a local minimizer of the energy (1). We note that many subspaces satisfy this condition (in particular, w.o.p. d-subspaces spanned by randomly sampled d vectors). Therefore, l_p minimization for multiple subspaces with $0 < p < 1$ will often lead to plenty of local minima.

This wealth of local minima clearly does not occur when $p = 1$ (or $p \geq 1$). It will be interesting, though difficult, to carefully analyze the number and depth of local minima for $p \geq 1$.

5.6. *The case of affine subspaces.* Our analysis was restricted to linear subspaces, though we believe that it can be extended to affine subspaces. Indeed, we can consider the affine Grassmannian [15], which distinguishes between subspaces according to both their offsets with respect to the origin (i.e., distances to closest linear subspaces of the same dimension) and their orientations (based on principal angles of the shifted linear subspaces). By assuming only affine subspaces intersecting a fixed ball, we can have a compact space. We can also generalize (70) (with a different function ψ_{μ_1}) and the estimates on δ_0 and κ_0 in Section 4.5.5 to the case of affine subspaces. We remark, though, that it is not obvious whether the metric on the affine Grassmannian is relevant for our applications, since it mixes two different quantities of different units (i.e., offset values and orientations) so that one can arbitrarily weigh their contributions. Also, the common strategy of using homogenous coordinates which transform d-dimensional affine subspaces in $\mathbb{R}^D$ to $(d+1)$-dimensional linear subspaces in $\mathbb{R}^{D+1}$ is not useful to us since it distorts the structure of both noise and outliers.

The minimization of the energy (1) over affine subspaces seems to result in more local minima than in the linear case, which can partially explain why numerical heuristics for minimizing (1) do not perform as well with affine subspaces as they do with linear ones. We are interested in further explanation of this phenomenon.

5.7. *The case of mixed dimensions.* It will be interesting to try to extend our analysis to linear subspaces of mixed dimensions $d_1, \dots, d_K$, known in advance. We believe that it is possible to extend Theorem 1.1 and its proof to this case. For this purpose, we suggest using the same distance for subspaces of the same dimension and defining the distance $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2)$ between linear subspaces L_1 and L_2 of different dimensions (with some abuse of notation) as follows: if $\dim(\mathrm{L}_1) < \dim(\mathrm{L}_2)$, then $\mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L}_2) = \min_{\mathrm{L} \in \mathrm{L}_2, \dim(\mathrm{L}) = \dim(\mathrm{L}_1)} \mathrm{dist}_{\mathrm{G}}(\mathrm{L}_1, \mathrm{L})$.

5.8. *Further performance guarantees for l_p-based HLM algorithms.* We are interested in extending our theory to analyze heuristics (like the K-subspaces) which try to minimize the l_p energy of (1) in practice.

5.9. *Asymptotic rates of convergence and sample complexity.* In Section 3.2 we demonstrated simple instances when noise is present and one cannot asymptotically recover the underlying subspaces by l_p minimization for all $p > 0$. One may still inquire about the existence of asymptotic limit different than the underlying subspaces and quantify the rate of convergence (depending on the mixture model parameters) to that limit. That is, assume that $\{\hat{\mathrm{L}}_1, \hat{\mathrm{L}}_2\}$ is the minimizer of $\mathbb{E}_\mu(l_p(\mathbf{x}, \mathrm{L}_1, \mathrm{L}_2))$ and $\{\hat{\mathrm{L}}_1^N, \hat{\mathrm{L}}_2^N\}$ is the minimizer of $\mathbb{E}_{\mu_N}(l_p(\mathbf{x}, \mathrm{L}_1, \mathrm{L}_2))$, where μ_N is an empirical distribution of i.i.d. sample of N points from μ. We first ask whether $\mathrm{dist}(\{\hat{\mathrm{L}}_1, \hat{\mathrm{L}}_2\}, \{\hat{\mathrm{L}}_1^N, \hat{\mathrm{L}}_2^N\}) \to 0$

as $N \to \infty$. If true, then we ask about the asymptotic rates of convergence. This will then allow a definition of a sample complexity for multiple subspaces as the number of samples required to achieve a prediction error within ε of the exact recovery of the K d-subspaces.

APPENDIX: SUPPLEMENTARY DETAILS

A.1. Proof of Lemma 2.1. We will use the following inequality for any $1 \le j \le K$, which is proved in [12], Section A.1.1:

$$\begin{aligned}&\mu_1(\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1^* : \operatorname{dist}(\mathbf{x}, \hat{\mathrm{L}}_j) < \beta \operatorname{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_j))\\ &\quad \le \psi_{\mu_1}\left(\frac{\pi\sqrt{d}}{2}\beta\right) \qquad \forall \beta > 0.\end{aligned} \tag{70}$$

We denote $\beta_1 = \frac{2}{\pi\sqrt{d}}\psi_{\mu_1}^{-1}(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K})$ (the existence of $\psi_{\mu_1}^{-1}(\frac{1+(2K-1)\mu_1(\{\mathbf{0}\})}{2K})$ follows the same proof as in [12], Section A.1.1) and combine (70) with the fact that $\operatorname{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_j) \ge \varepsilon$ for any $1 \le j \le K$ to obtain that

$$\begin{aligned}&\mu_1(\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1^* \setminus \{\mathbf{0}\} : \operatorname{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) < \beta_1 \varepsilon)\\ &\qquad = \mu_1(\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1^* \setminus \{\mathbf{0}\} : \operatorname{dist}(\mathbf{x}, \hat{\mathrm{L}}_1) < \beta_1 \operatorname{dist}_{\mathrm{G}}(\mathrm{L}_1^*, \hat{\mathrm{L}}_1))\\ &\qquad \le \frac{1 + (2K-1)\mu_1(\{\mathbf{0}\})}{2K} - \mu(\{\mathbf{0}\})\\ &\qquad = \frac{1 - \mu_1(\{\mathbf{0}\})}{2K}.\end{aligned}$$

Consequently,

$$\begin{aligned}&\mu_1\left(\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1^* : \operatorname{dist}\left(\mathbf{x}, \bigcup_{j=1}^{K} \hat{\mathrm{L}}_1\right) \ge \beta_1 \varepsilon\right)\\ &\qquad \ge 1 - \mu(\{\mathbf{0}\}) - \sum_{i=1}^{K} \mu_1(\mathbf{x} \in \mathrm{B}(\mathbf{0}, 1) \cap \mathrm{L}_1^* \setminus \{\mathbf{0}\} : \operatorname{dist}(\mathbf{x}, \hat{\mathrm{L}}_i) < \beta_1 \varepsilon)\\ &\qquad \ge (1 - \mu_1(\{\mathbf{0}\}))/2,\end{aligned}$$

and, thus, by Chebyshev's inequality the lemma is concluded as follows:

$$\begin{aligned}&\mathbb{E}_{\mu_1}(e_{l_p}(\mathbf{x}, \hat{\mathrm{L}}_1)) \ge \beta_1^p \varepsilon^p / 2\\ &\qquad\qquad = \frac{(1 - \mu_1(\{\mathbf{0}\}))2^{p-1}\psi_{\mu_1}^{-1}((1 + (2K-1)\mu_1(\{\mathbf{0}\}))/(2K))^p \varepsilon^p}{(\pi\sqrt{d})^p}\\ &\qquad\qquad = \tau_0 \varepsilon^p.\end{aligned}$$

A.2. Proof of Proposition 3.1. The proof is an immediate consequence of the following inequality, which uses an arbitrary $\mathrm{L}_1 \in \mathrm{G}(D, d)$ and the notation $\mathrm{Y}'_i = \mathrm{Y}_i(\mathrm{L}'_1, \ldots, \mathrm{L}'_K)$, $1 \leq i \leq K$:

$$\begin{aligned} 0 &\leq \mathbb{E}_\nu(e_{l_p}(\mathbf{x}, \mathrm{L}_1, \mathrm{L}'_2, \ldots, \mathrm{L}'_K)) - \mathbb{E}_\nu(e_{l_p}(\mathbf{x}, \mathrm{L}'_1, \ldots, \mathrm{L}'_K)) \\ &\leq \mathbb{E}_\nu(I(\mathbf{x} \in \mathrm{Y}'_1) e_{l_p}(\mathbf{x}, \mathrm{L}_1)) + \sum_{2 \leq i \leq K} \mathbb{E}_\nu(I(\mathbf{x} \in \mathrm{Y}'_i) e_{l_p}(\mathbf{x}, \mathrm{L}'_i)) \\ &\quad - \sum_{1 \leq i \leq K} \mathbb{E}_\nu(I(\mathbf{x} \in \mathrm{Y}'_i) e_{l_p}(\mathbf{x}, \mathrm{L}'_i)) \\ &= \mathbb{E}_\nu(I(\mathbf{x} \in \mathrm{Y}'_1) e_{l_p}(\mathbf{x}, \mathrm{L}_1)) - \mathbb{E}_\nu(I(\mathbf{x} \in \mathrm{Y}'_1) e_{l_p}(\mathbf{x}, \mathrm{L}'_1)). \end{aligned}$$

A.3. Proof of Lemma 4.2: Geometric sensitivity.

We will first show that there exists $\mathbf{x}_0 \in \mathrm{B}(\mathbf{0}, 1)$ such that

$$\operatorname{dist}(\mathbf{x}_0, \mathrm{L}^*_1) = \operatorname{dist}(\mathbf{x}_0, \mathrm{L}^*_2) < \min_{3 \leq i \leq K} \operatorname{dist}(\mathbf{x}_0, \mathrm{L}^*_i). \tag{71}$$

We verify (71) in two cases: $d^* = d$ and $d^* = D - d$. We will then prove that (71) implies (47). Throughout the proof we denote the principal vectors of L^*_2 and L^*_1 by $\{\hat{\mathbf{v}}_i\}_{i=1}^{d^*}$ and $\{\mathbf{v}_i\}_{i=1}^{d^*}$, respectively.

A.3.1. *Part* I*: Proof of (71) when $d^* = d$.* We define

$$\mathbf{x}_0 = (\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*}) / \|\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*}\|$$

and arbitrarily fix $i_0 > 3$ and $\mathbf{v}_0 \in \mathrm{L}^*_{i_0}$. We will show that

$$\operatorname{ang}(\mathbf{x}_0, \mathbf{v}_0) > \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 \tag{72}$$

and consequently conclude (71) as follows:

$$\begin{aligned} \operatorname{dist}(\mathbf{x}_0, \mathrm{L}^*_{i_0}) &\geq \sin(\operatorname{ang}(\mathbf{x}_0, \mathbf{v}_0)) > \sin(\theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2) = \operatorname{dist}(\mathbf{x}_0, \mathrm{L}^*_1) \\ &= \operatorname{dist}(\mathbf{x}_0, \mathrm{L}^*_2). \end{aligned}$$

We can easily verify a weaker version of (72) where the inequality is not necessarily strict. Indeed, using elementary geometric estimates and the fact that the intersections of the d-subspaces $\{\mathrm{L}^*_i\}_{i=1}^K$ are empty [which follows from (45)], we obtain that

$$\begin{aligned} \operatorname{ang}(\mathbf{x}_0, \mathbf{v}_0) &\geq \operatorname{ang}(\mathbf{v}_{d^*}, \mathbf{v}_0) - \operatorname{ang}(\mathbf{v}_{d^*}, \mathbf{x}_0) \geq \theta_{d^*}(\mathrm{L}^*_{i_0}, \mathrm{L}^*_1) - \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 \\ &\geq \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1) - \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 = \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2. \end{aligned} \tag{73}$$

At last, we show that (73) cannot be an equality. Indeed, if the first inequality in (73) is an equality, then $\mathbf{v}_0$, $\mathbf{v}_{d^*}$ and $\mathbf{x}_0$ are on a geodesic line within the sphere S^{D-1}. Combining this with the assumption that all other inequalities in (73) are equalities, we obtain that $\operatorname{ang}(\mathbf{x}_0, \mathbf{v}_0) = \theta_{d^*}(\mathrm{L}^*_2, \mathrm{L}^*_1)/2 = \operatorname{ang}(\mathbf{x}_0, \mathbf{v}_{d^*}) = \operatorname{ang}(\mathbf{x}_0, \hat{\mathbf{v}}_{d^*})$. This implies that either $\mathbf{v}_0 = \hat{\mathbf{v}}_{d^*}$ or $\mathbf{v}_0 = \mathbf{v}_{d^*}$, which contradicts (45).

A.3.2. *Part* II*: Proof of (71) when $d^* = D - d$.* It follows from basic dimension equalities of subspaces and (45) that for all $2 \leq i \leq K : \dim(\mathrm{L}^*_1 \cup$

$\mathrm{L}_i^*) = D$ and $\dim(\mathrm{L}_1^* \cap \mathrm{L}_i^*) = 2d - D$. We denote by K_0 the integer in $\{0, \ldots, K\}$ such that for any $3 \leq i \leq K_0 : \mathrm{L}_1^* \cap \mathrm{L}_i^* = \mathrm{L}_1^* \cap \mathrm{L}_2^*$ and for any $i > K_0 : \mathrm{L}_1^* \cap \mathrm{L}_i^* \neq \mathrm{L}_1^* \cap \mathrm{L}_2^*$ (the existence of K_0 may require reordering of the indices of the subspaces $\{\mathrm{L}_i^*\}_{i=3}^K$). In order to define $\mathbf{x}_0$ in the current case, we let $\mathbf{x}_1 = (\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*})/\|\hat{\mathbf{v}}_{d^*} + \mathbf{v}_{d^*}\|$, $\mathbf{x}_2$ be an arbitrarily fixed unit vector in $\mathrm{L}_1^* \cap (\mathrm{L}_2^* \setminus \bigcup_{K_0 < i \leq K} \mathrm{L}_i^*)$, $\varepsilon_0 = \operatorname{dist}(\mathbf{x}_2, \bigcup_{K_0 < i \leq K} \mathrm{L}_i^*)$ and

$$\mathbf{x}_0 = \mathbf{x}_2/2 + \varepsilon_0 \mathbf{x}_1/5.$$

We first claim that

$$\operatorname{dist}(\mathbf{x}_0, \mathrm{L}_1^*) = \operatorname{dist}(\mathbf{x}_0, \mathrm{L}_2^*) < \min_{3 \leq j \leq K_0} \operatorname{dist}(\mathbf{x}_0, \mathrm{L}_j^*). \tag{74}$$

Indeed, we can remove $\mathrm{L}_1^* \cap \mathrm{L}_2^*$ from the subspaces $\{\mathrm{L}_i^*\}_{i=1}^{K_0}$ and obtain subspaces of dimension $D - d$ intersecting each other at the origin. We can then rewrite (74) by replacing $\{\mathrm{L}_i^*\}_{i=1}^{K_0}$ with their reduced version and $\mathbf{x}_0$ with $\mathbf{x}_1$. The argument of Section A.3.1 thus proves this equation.

We conclude (71) by combining (74) with the following observation:

$$\begin{aligned}
\operatorname{dist}(\mathbf{x}_0, \mathrm{L}_1^*) &= \varepsilon_0 \operatorname{dist}(\mathbf{x}_1, \mathrm{L}_1^*)/5 \leq \varepsilon_0/5 < \operatorname{dist}\left(\mathbf{x}_2/2, \bigcup_{K_0 < j \leq K} \mathrm{L}_j^*\right) - \varepsilon_0/5 \\
&\leq \operatorname{dist}\left(\mathbf{x}_2/2 + \varepsilon_0 \mathbf{x}_1/5, \bigcup_{K_0 < j \leq K} \mathrm{L}_j^*\right) = \min_{K_0 < i \leq K} \operatorname{dist}(\mathbf{x}_0, \mathrm{L}_i^*).
\end{aligned} \tag{75}$$

A.3.3. *Part* III*: Deriving (47) from (71) in a simple case.* We note that (71) implies that

$$\mathbf{x}_0 \in (\mathrm{Y}_1 \cup \mathrm{Y}_2 \cup (\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2)) \cap (\hat{\mathrm{Y}}_1 \cup \hat{\mathrm{Y}}_2 \cup (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2)) \tag{76}$$

and, consequently,

$$\mathrm{B}(\mathbf{x}_0, \varepsilon) \subset (\mathrm{Y}_1 \cup \mathrm{Y}_2 \cup (\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2)) \cap (\hat{\mathrm{Y}}_1 \cup \hat{\mathrm{Y}}_2 \cup (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2)). \tag{77}$$

We will deduce here (47) from (77) in the simpler case: $\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \varepsilon) \neq \bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2 \cap \mathrm{B}(\mathbf{x}_0, \varepsilon)$.

Using (77) and the fact that $\mathcal{L}_D(\bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2) = 0$, we may choose $\mathbf{y} \in (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \varepsilon)) \cap (\mathrm{Y}_1 \cup \mathrm{Y}_2)$; WLOG we assume instead of the latter condition that $\mathbf{y} \in (\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \varepsilon)) \cap \mathrm{Y}_1$. By slightly perturbing $\mathbf{y}$ we can choose another point $\mathbf{y}_0$ such that $\mathbf{y}_0 \in \hat{\mathrm{Y}}_2$ and $\mathbf{y}_0 \in \mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1$. It follows from the continuity of the distance function that there exists a small $\eta > 0$ such that $(\hat{\mathrm{Y}}_1 \setminus \mathrm{Y}_1) \cup (\mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1) \supseteq \mathrm{Y}_1 \setminus \hat{\mathrm{Y}}_1 \supset \mathrm{B}(\mathbf{y}_0, \eta)$, which proves (47).

A.3.4. *Part* IV*: Deriving (47) from (71) in the complementary case.* At last, we assume that $\bar{\hat{\mathrm{Y}}}_1 \cap \bar{\hat{\mathrm{Y}}}_2 \cap \mathrm{B}(\mathbf{x}_0, \varepsilon) = \bar{\mathrm{Y}}_1 \cap \bar{\mathrm{Y}}_2 \cap \mathrm{B}(\mathbf{x}_0, \varepsilon)$. We show here that it leads to the contradiction: $\hat{\mathrm{L}}_2 = \mathrm{L}_2^*$.

We note that the sets of solutions in $\mathrm{B}(\mathbf{x}_0,\varepsilon)$ of the equations $\mathbf{x}^T(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})\mathbf{x}=0$ and $\mathbf{x}^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})\mathbf{x}=0$ are $\bar{\hat{\mathrm{Y}}}_1\cap\bar{\hat{\mathrm{Y}}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon)$ and $\bar{\mathrm{Y}}_1\cap\bar{\mathrm{Y}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon)$, respectively. In view of (77), these solution sets coincide. They are $(D-1)$-manifolds and, thus, their $(D-1)$-dimensional tangent spaces at $\mathbf{x}_0$, that is, $\mathbf{x}_0^T(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})=\mathbf{0}$ and $\mathbf{x}_0^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})=\mathbf{0}$, also coincide. Consequently, we have that $\mathbf{x}_0^T(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})=t_0\mathbf{x}_0^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})$ for some $t_0\neq 0$. Similarly, for any $\mathbf{x}_1\in\bar{\hat{\mathrm{Y}}}_1\cap\bar{\hat{\mathrm{Y}}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon)$, we have $\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})=t_1\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})$ for some $t_1\neq 0$. We note that $t_1=t_0$ by the following argument: $t_1\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})\mathbf{x}_0=\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})\mathbf{x}_0=t_0\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})\mathbf{x}_0$. Therefore, there exists $t\neq 0$ such that for any $\mathbf{x}_1\in\bar{\hat{\mathrm{Y}}}_1\cap\bar{\hat{\mathrm{Y}}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon)$,

$$\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})=t\mathbf{x}_1^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2}). \tag{78}$$

Since the tangent space of $\bar{\hat{\mathrm{Y}}}_1\cap\bar{\hat{\mathrm{Y}}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon)$ [or, equivalently, $\mathbf{x}^T(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})\mathbf{x}=0$] at $\mathbf{x}_0$ has dimension $D-1$, the subspace $\mathrm{L}_0^*=\mathrm{Sp}(\bar{\hat{\mathrm{Y}}}_1\cap\bar{\hat{\mathrm{Y}}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon))$ [i.e., the closure of all finite linear combinations of vectors in $\bar{\hat{\mathrm{Y}}}_1\cap\bar{\hat{\mathrm{Y}}}_2\cap\mathrm{B}(\mathbf{x}_0,\varepsilon)$] has dimension at least $D-1$. In view of (78), L_0^* satisfies

$$P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})=tP_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2}). \tag{79}$$

Due to the symmetry of $(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})$ and $(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})$, we have the following equivalent formulation of (79):

$$(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^*}=(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^*}. \tag{80}$$

Furthermore, using the fact that $(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})$ and $(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})$ have trace 0, we obtain that

$$\begin{aligned}\operatorname{tr}(P_{\mathrm{L}_1^{*\perp}}(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^{*\perp}})&=-\operatorname{tr}(P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^*})\\&=-t\cdot\operatorname{tr}(P_{\mathrm{L}_0^*}(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^*})\\&=t\cdot\operatorname{tr}(P_{\mathrm{L}_0^{*\perp}}(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^{*\perp}}).\end{aligned} \tag{81}$$

Since $P_{\mathrm{L}_0^{*\perp}}$ is at most one-dimensional, (81) can be rewritten as

$$P_{\mathrm{L}_0^{*\perp}}(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})P_{\mathrm{L}_0^{*\perp}}=t\cdot(P_{\mathrm{L}_0^{*\perp}}(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})P_{\mathrm{L}_0^{*\perp}}). \tag{82}$$

Combining (79), (80) and (82), we obtain that $(P_{\mathrm{L}_1^*}-P_{\hat{\mathrm{L}}_2})=t(P_{\mathrm{L}_1^*}-P_{\mathrm{L}_2^*})$, equivalently,

$$P_{\hat{\mathrm{L}}_2}=(1-t)P_{\mathrm{L}_1^*}+tP_{\mathrm{L}_2^*}. \tag{83}$$

We conclude the desired contradiction in two different cases. Assume first that $t<1$ and let $\mathbf{v}_0$ be an arbitrary unit vector in L_2^*. We note that

$\mathbf{v}_0^T P_{\hat{\mathrm{L}}_2} \mathbf{v}_0 = 1$ as well as $(1-t)\mathbf{v}_0^T P_{\mathrm{L}_1^*} \mathbf{v}_0 = 1 - t\mathbf{v}_0^T P_{\mathrm{L}_2^*} \mathbf{v}_0 \geq 1-t$. Consequently, $\mathbf{v}_0^T P_{\mathrm{L}_1^*} \mathbf{v}_0 = 1$, that is, $\mathbf{v}_0 \in \mathrm{L}_1^*$ and, thus, we obtain the following contradiction with (45): $\mathrm{L}_1^* = \hat{\mathrm{L}}_2$ [in view of (83), this is equivalent with $\hat{\mathrm{L}}_2 = \mathrm{L}_2^*$]. Next, assume that $t \geq 1$ and, as before, $\mathbf{v}_0$ is an arbitrary unit vector in $\mathrm{L}_2^{*\perp}$. In this case, $\mathbf{v}_0^T P_{\hat{\mathrm{L}}_2} \mathbf{v}_0 = (1-t)\mathbf{v}_0^T P_{\mathrm{L}_1^*} \mathbf{v}_0 + t\mathbf{v}_0^T P_{\mathrm{L}_2^*} \mathbf{v}_0 \leq 0 + 0 = 0$. Therefore, $\mathbf{v}_0 \in \hat{\mathrm{L}}_2^{\perp}$ and we obtain the following contradiction with (45): $\mathrm{L}_2^* = \hat{\mathrm{L}}_2$. Equation (47) is thus proved.

Acknowledgments. Our collaboration with Arthur Szlam on efficient and fast algorithms for hybrid linear modeling (especially via geometric l_1 minimization) inspired this investigation. We thank John Wright for interesting discussions and J. Tyler Whitehouse for commenting on an earlier version of this manuscript. Thanks to the Institute for Mathematics and its Applications (IMA), in particular, Doug Arnold and Fadil Santosa, for holding a workshop on multi-manifold modeling that G. Lerman co-organized and T. Zhang participated in. G. Lerman thanks David Donoho for inviting him for a visit to Stanford University in Fall 2003 and for stimulating discussions at that time on the intellectual responsibilities of mathematicians analyzing massive and high-dimensional data as well as general advice. Those discussions effected G. Lerman's research program and his mentorship (T. Zhang is a Ph.D. candidate advised by G. Lerman).

REFERENCES

[1] Aldroubi, A., Cabrelli, C. and Molter, U. (2008). Optimal non-linear models for sparsity and sampling. *J. Fourier Anal. Appl.* **14** 793–812. MR2461607

[2] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York. MR0771294

[3] Arias-Castro, E., Chen, G. and Lerman, G. (2011). Spectral clustering based on local linear approximations. *Electron. J. Statist.* **5** 1537–1587.

[4] Bendich, P., Wang, B. and Mukherjee, S. (2010). Towards stratification learning through homology inference. Available at http://arxiv.org/abs/1008.3572.

[5] Bradley, P. S. and Mangasarian, O. L. (2000). k-plane clustering. *J. Global Optim.* **16** 23–32. MR1770524

[6] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2009). Robust principal component analysis? Unpublished manuscript. Available at arXiv:0912.3599.

[7] Chen, G. and Lerman, G. (2009). Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Found. Comput. Math.* **9** 517–558. MR2534403

[8] Chen, G. and Lerman, G. (2009). Spectral curvature clustering (SCC). *Int. J. Comput. Vision* **81** 317–330.

[9] Costeira, J. and Kanade, T. (1998). A multibody factorization method for independently moving objects. *Int. J. Comput. Vis.* **29** 159–179.

[10] Ho, J., Yang, M., Lim, J., Lee, K. and Kriegman, D. (2003). Clustering appearances of objects under varying illumination conditions. In *Proceedings of International Conference on Computer Vision and Pattern Recognition* **1** 11–18. IEEE Computer Society, Madison, WI.

[11] Kanatani, K. (2001). Motion segmentation by subspace separation and model selection. In *Proc. of 8th ICCV* **3** 586–591. IEEE, Vancouver, Canada.

[12] Lerman, G. and Zhang, T. (2010). l_p-Recovery of the most significant subspace among multiple subspaces with outliers. Unpublished manuscript. Available at `http://arxiv.org/abs/1012.4116`.

[13] Ma, Y., Derksen, H., Hong, W. and Wright, J. (2007). Segmentation of multivariate mixed data via lossy coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29** 1546–1562.

[14] Ma, Y., Yang, A. Y., Derksen, H. and Fossum, R. (2008). Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Rev.* **50** 413–458. MR2429444

[15] Mattila, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability. Cambridge Studies in Advanced Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR1333890

[16] Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140. MR0600539

[17] Pollard, D. (1982). A central limit theorem for k-means clustering. *Ann. Probab.* **10** 919–926. MR0672292

[18] Shawe-Taylor, J., Williams, C. K. I., Cristianini, N. and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inform. Theory* **51** 2510–2522. MR2246374

[19] Tao, T. (2011). Topics in random matrix theory. Available at http://terrytao.files.wordpress.com/2011/02/matrix-book.pdf.

[20] Tipping, M. and Bishop, C. (1999). Mixtures of probabilistic principal component analysers. *Neural Comput.* **11** 443–482.

[21] Torr, P. H. S. (1998). Geometric motion segmentation and model selection. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **356** 1321–1340. MR1627069

[22] Tseng, P. (2000). Nearest q-flat to m points. *J. Optim. Theory Appl.* **105** 249–252. MR1757267

[23] Vidal, R., Ma, Y. and Sastry, S. (2005). Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.* **27** 1945–1959.

[24] Yan, J. and Pollefeys, M. (2006). A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate. In *ECCV* **4** 94–106.

[25] Zhang, T., Szlam, A. and Lerman, G. (2009). Median K-flats for hybrid linear modeling with many outliers. In *Computer Vision Workshops* (*ICCV Workshops*), *IEEE 12th International Conference on Computer Vision* 234–241. IEEE, Tokyo, Japan.

[26] Zhang, T., Szlam, A., Wang, Y. and Lerman, G. (2010). Hybrid linear modeling via local best-fit flats. Available at `http://arxiv.org/abs/1010.3460`.

[27] Zhang, T., Szlam, A., Wang, Y. and Lerman, G. (2010). Randomized hybrid linear modeling by local best-fit flats. In *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 1927–1934. IEEE, San Francisco, CA.

Department of Mathematics
University of Minnesota
127 Vincent Hall
206 Church Street SE
Minneapolis, Minnesota 55455
USA
E-mail: lerman@umn.edu
zhang620@umn.edu